THE CAMBRIDGE HANDBOOK OF RESEARCH METHODS AND STATISTICS FOR THE SOCIAL AND BEHAVIORAL SCIENCES

Edited by Austin Lee Nichols and John Edlund

Volume 1: Building a Program of Research

The Cambridge Handbook of Research Methods and Statistics for the Social and Behavioral Sciences

Volume 1

The first of three volumes, this book, covers a variety of issues important in developing, designing, and analyzing data to produce high-quality research efforts and cultivate a productive research career. First, leading scholars from around the world provide a step-by-step guide to doing research in the social and behavioral sciences. After discussing some of the basics, the various authors next focus on the important building blocks of any study. In Part III, various types of quantitative and qualitative research designs are discussed, and advice is provided regarding best practices of each. The volume then provides an introduction to a variety of important and cutting-edge statistical analyses. In the last part of the volume, nine chapters provide information related to what it takes to have a long and successful research career. Throughout the book, examples and real-world research efforts from dozens of different disciplines are discussed.

AUSTIN LEE NICHOLS is Associate Professor of Organizational Psychology at Central European University in Vienna, Austria. Prior to his current position, he worked in various faculty and research positions around the world in both psychology and management. He has published in journals across a variety of research disciplines and has won awards for his teaching, research, and service from various global institutions.

JOHN E. EDLUND is Professor of Psychology at the Rochester Institute of Technology, USA, and serves as the Research Director of Psi Chi: The International Honor Society in Psychology. He has won numerous awards related to teaching and is passionate about the improvement of research methods and the dissemination of psychological knowledge to the world.

Cambridge Handbooks in Psychology

The Cambridge Handbook of Research Methods and Statistics for the Social and Behavioral Sciences

Volume 1: Building a Program of Research

Edited by

Austin Lee Nichols
Central European University

John E. Edlund
Rochester Institute of Technology

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment, a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781316518526

DOI: 10.1017/9781009010054

First published 2023

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

A Cataloging-in-Publication data record for this book is available from the Library of Congress.

ISBN 978-1-316-51852-6 Hardback
ISBN 978-1-108-99524-5 Paperback

Cambridge University Press & Assessment has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

Most importantly, I dedicate this book to my wife, family, and friends who put up with me reading and editing chapters at some very weird times and days. I would also like to thank the numerous coffee shops, mostly in Vienna and Orlando, for providing the energy and focus to edit the many wonderful chapters contained in this book.

Austin Lee Nichols

I dedicate this book to my awesome wife and children. Without you, all that I have and do would not be possible. You three are my everything.

John E. Edlund

Contents

Figures

Tables

Contributors

AARON R. SEITZ, University of California, Riverside

ALEXIS B. AVERY, University of Wisconsin–Madison

ANNE MOYER, Stony Brook University

ANTHONY J. GAMBINO, University of Connecticut

C. SHAWN GREEN, University of Wisconsin–Madison

CHARLES S. REICHARDT, University of Denver

CHRISTIAN S. CRANDALL, University of Kansas

CHRISTIAN UNKELBACH, University of Cologne

CRAIG A. ANDERSON, Iowa State University

D. BETSY MCCOACH, University of Connecticut

DAMON ABRAHAM, University of Denver

DANIEL P. CORTS, Augustana College

DANIEL STORAGE, University of Denver

DAVID GORETZKO, Ludwig-Maximilians-Universität München

DAVID S. FESTINGER, Philadelphia College of Osteopathic Medicine

DOLORES ALBARRACIN, University of Pennsylvania

EDGAR ERDFELDER, University of Mannheim

ELISABETTA RUSPINI, University of Milano-Bicocca

ELIZABETH COLLINS, University of Stirling

GINETTE BLACKHART, East Tennessee State University

GLYNIS M. BREAKWELL, University of Bath

HANNA K. PILLION, US Customs and Border Protection

HANNAH R. CALLAHAN, Philadelphia College of Osteopathic Medicine

HOWARD C. NUSBAUM, University of Chicago

IGNACIO FERRERO, University of Navarra

JASON MILLER, Michigan State University

JAVIER PINTO, University of Los Andes

JEFF GREENBERG, University of Arizona

JEFFREY M. CUCINA, US Customs and Border Protection

JENNIFER N. BAUMGARTNER, University of California, San Diego

JESSE CHANDLER, Mathematica; University of Michigan

JESSICA GUREVITCH, Stony Brook University

JOCELYN PARONG, University of Wisconsin–Madison

JOHN F. DOVIDIO, Yale University

JONATHAN A. MUIR, Emory University

JORDAN R. WAGGE, Avila University

KAREN L. DUGOSH, Public Health Management Corporation

KATHLEEN O'SULLIVAN, University College Cork

KAYONNE CHRISTY, University of British Columbia

KELLY CUCCOLO, Alma College

KEVIN A. BYLE, US Customs and Border Protection

KEVIN B. WRIGHT, George Mason University

KLAUS FIEDLER, Heidelberg University

LAITH AL-SHAWAF, University of Colorado, Colorado Springs

LISA L. HARLOW, University of Rhode Island

MANINDER SINGH SETIA, MGM Institute of Health Sciences

MARGARET DENNY, University of Maribor

MARIYA VODYANYK, University of California, Irvine

MARK SCHALLER, University of British Columbia

MARTHA S. ZLOKOVICH, Psi Chi, The International Honor Society in Psychology

MARTIN SCHNUERCH, University of Mannheim

MARTIN SELLBOM, University of Otago

MARY G. CAREY, University of Rochester Medical Center

MARY MOUSSA ROGERS, University of South Carolina Aiken

NICKY HAYES, Professional Development Foundation

RACHEL A. HOUGH, Public Health Management Corporation

RACHEL ADAMS GOERTEL, Roberts Wesleyan College, Rochester

REX B. KLINE, Concordia University, Montréal, Canada

ROGER WATT, University of Stirling

SARAH D. NEWTON, University of Connecticut

SHELDON SOLOMON, Skidmore College

SICONG LIU, University of Illinois at Urbana–Champaign

SINIKKA ELLIOTT, University of British Columbia

SIQI XIAO, University of British Columbia

SOLVEIG A. CUNNINGHAM, Emory University

SUSANNE M. JAEGGI, University of California, Irvine

SUZANNE DENIEFFE, Waterford Institute of Technology

TAMERA R. SCHNEIDER, Baruch College – CUNY

THOMAS F. DENSON, University of New South Wales

TODD K. SHACKELFORD, Oakland University

TOM PYSZCZYNSKI, University of Colorado at Colorado Springs

TRAVIS D. CLARK, University of North Dakota

WENDY M. BRUNNER, Bassett Medical Center

YURI JADOTTE, Stony Brook University

YZAR S. WEHBE, Oakland University

ZOLTAN DIENES, University of Sussex

Preface

The *Cambridge Handbook of Research Methods and Statistics for the Social and Behavioral Sciences* is meant to be the most comprehensive and contemporary collection of topics related to research methods and statistics spanning these related yet extremely diverse fields of research. This first volume, *Building a Program of Research*, provides researchers at all levels a starting point along with the tools to build a successful research career in one of these fields. Although each chapter provides a substantial contribution to this end, together the individual chapters combine to provide the knowledge needed to be a successful researcher in the social and behavioral sciences.

Throughout these chapters, the leading researchers in a variety of disciplines seek to share their knowledge and experience in a way that is both accessible and useful. They do so by writing in a way that is understandable to novice researchers and also deeply discusses the challenges related to each topic and provides new information to highly experienced scientists. This volume begins with issues related to building theory and generating promising ideas, includes detailed topics related to each of the steps involved in the research process, and provides ethical considerations that should be at the forefront of any research project.

Volume 1 next focuses on detailed building blocks of any research endeavor, including issues related to recruitment of participants, providing informed consent, awareness of and amelioration of experimenter effects, and how best to debrief and probe participants at the conclusion of the study. The chapters that follow get into the nitty gritty of data collection by focusing on, giving examples of, and providing advice for a variety of study designs and methodological approaches. Subsequently, the experts address several considerations for analyzing a variety of quantitative and qualitative data, ranging from cleaning the data to running descriptive statistics to introducing higher-level modeling techniques.

The volume finishes by providing real-world advice, from extremely successful researchers, that will help even the most experienced scientists to further their career. Topics include designing a line of research, publishing and presenting one’s research, successfully collaborating, handling and reviewing your own and others’ research submissions, grant writing, teaching methods and statistics, and even options and applications for researchers outside of a traditional academic context. In all, the authors in this volume span over a dozen disciplines, many more countries, and have amassed successful research careers leading to numerous publications and acknowledgments. It is for this reason that we are confident in their ability to teach you and to help you progress in your career as a scientist.

PART I

From Idea to Reality: The Basics of Research

1 Promises and Pitfalls of Theory

Yzar S. Wehbe, Todd K. Shackelford, and Laith Al-Shawaf

Abstract

We present an overview of the role, benefits, and drawbacks of theory in scientific research, particularly in the social and behavioral sciences. We discuss what theory is and what it is not. We also focus on some key elements of theory such as its ability to explain phenomena at multiple parallel levels of analysis. Evolutionary theory is offered as an example that illustrates the importance of conceptual integration across different disciplines. We further describe the key characteristics of good theories, such as parsimony, depth, breadth, and coherence (both internal and external), and we encourage the use of "coherence stress-tests" to help refine theory. We then discuss 4 advantages and 10 disadvantages of using theory in social and behavioral science research. Finally, we suggest conceptual tools and provide a list of recommendations for theory-driven research. We hope this chapter will help in the complex pursuit of improving research practices in the social and behavioral sciences.

Keywords: Top-down Approach, Theory Building, Conceptual Integration, Levels of Analysis, Parsimony, Interdisciplinarity, Evolutionary Theory

> *One of the strengths of scientific inquiry is that it can progress with any mixture of empiricism, intuition, and formal theory that suits the convenience of the investigator. Many sciences develop for a time as exercises in description and empirical generalization. Only later do they acquire reasoned connections within themselves and with other branches of knowledge.*
>
> (Williams, 1966, p. 20)

Introduction

The goal of science is to understand the world. This is much easier to do when we develop and rely on good theories (Goetz & Shackelford, 2006). Strong theoretical foundations help a researcher make predictions, ask the right questions, and interpret data in a meaningful way. Research lacking theory is, in a sense, exploratory, meaning that it is consigned to trial and error – an inefficient way of accumulating knowledge. Some scholars in the social and behavioral sciences have even contended that empirical findings generated atheoretically are less convincing and thus less likely to be used in practical applications (e.g., Burns, 2011). However, as we discuss later in the chapter, there are also ways in which theory can lead us astray.

To oversimplify, a scientific theory is a set of ideas that has the power to explain and predict real phenomena, albeit never fully or perfectly. For social and behavioral scientists, a strong grounding in theory is our best hope for understanding human cognition and behavior. In science, theory is generated, developed, amended, and replaced on evidentiary grounds. But how do we generate, develop, and amend theories? And how do we know which theories are fruitful and which may be leading us astray?

Many scholars have lamented the overuse, underuse, and misuse of theory in the social sciences (Borsboom et al., 2020; Fried, 2020; Gigerenzer, 2009, 2010; Meehl, 1978; Muthukrishna & Henrich, 2019; Symons, 1992; Tooby & Cosmides, 2015). This chapter describes some of the benefits and dangers of theorizing in the social sciences and offers recommendations for developing and evaluating theory. Theory can inspire and guide research, but what counts as theory, and what does not?

What Theory Is and What It Is Not

Much empirical research focuses not on *explaining* phenomena (making causal claims about how a phenomenon came to be) but simply on *describing* phenomena. For the purposes of this chapter, theory can be distinguished from descriptive research in that theory does not only describe facts, but theory also makes causal and explanatory claims about the world. By contrast, examples of descriptive research may include generalizations, regularities, typologies, and taxonomies. Such empirical research offers descriptions of the world, but does not offer causal explanations (indeed, empirical generalizations often *require* explanations). However, empirical generalizations that offer no explanations have sometimes been erroneously labeled “theory,” probably because they offer some predictive utility. Nettle (2021) illustrates this loose usage of the term “theory” with reference to “social identity theory.” However, social identity theory does not make causal claims; it only describes and predicts humans’ interest in their social identities. Theories should go beyond describing and predicting and afford a path to understanding. Empirical generalizations tell us about phenomena in the world, including their antecedents and consequents, but only theory can *explain* these effects, accounting for why they are the way they are or why we do not see different phenomena instead.

A common description of theory is a nomological network, namely, a representation of relationships between well-defined constructs (Cronbach & Meehl, 1955). Due to this definition, questions of theory underlie questions about validity because these also involve mapping the relationships between constructs. Causal links between constructs that describe real-world phenomena are key to questions about various forms of evidence for validity, including content and response processes (e.g., does our measure accurately capture all the aspects of a phenomenon?). To assess if a measure is accurate in detecting a phenomenon, we attempt to determine the measure’s criterion validity. To do so, we need to have a well-specified theory about when, why, how, and in what contexts the phenomenon in question will affect and be affected by other phenomena (Borsboom et al., 2004). A robust theoretical grounding, then, is key to validity (Gray, 2017).

Some scholars have likened constructing a theory to erecting a building using uneven bricks, whereby each brick is a study or a fact. Gray (2017) uses another metaphor for theory to remind us that it is not enough to focus on research methods alone: “The quest for reliable research methods – for making good bricks – is certainly noble, but the mere collection of reliable studies does not make for good science. We must remember that we scientists are not only brickmakers but also architects; we need to turn our attention back to building – to theory” (p. 732). The key point is that it is much easier to assemble a collection of uneven bricks into a robust and useful building when a good blueprint is available (Poincaré, 1905). Theories are like blueprints that help us understand how the empirical generalizations we discover fit with each other like pieces in a puzzle. Descriptions of some key concepts in theoretical and descriptive research can be found in Table 1.1. These concepts do not always have clear boundaries. For example, at what point does a theory become a paradigm? Nor do concepts have universally agreed definitions and usages. For example, “principle” is sometimes used to describe both theoretical tenets and empirical regularities. As a result, this table should be considered a rough

Table 1.1 *Definitions and descriptions of common theoretical and descriptive entities used in research*

Theoretical terms	Definitions and descriptions
Paradigm	A cumulative integrative theoretical framework. A collection of general ways of viewing the world, typically composed of interwoven theoretical claims and necessary auxiliary assumptions. A theory that is broad enough to guide an entire field of study is often referred to as a paradigm.
Theory	A set of ideas for explaining and predicting phenomena in the world. A proposition about the suspected relationship between variables. It is broader than a hypothesis and may be used to generate specific hypotheses. This typically explains a broad range of phenomena.
Causal hypothesis	A proposed explanatory link between two constructs. It is more specific than a theory and broader than a prediction.
Prediction	A testable proposition that is derived from, or generated on the basis of, a causal hypothesis. Hypotheses are tested via the specific predictions they yield.
Descriptive terms	**Definitions and descriptions**
Law, rule, or principle	Empirical generalizations that successfully describe an observed regularity. They are not explanations but are expected to be *explainable* (i.e., we can hope to use theory to explain why these generalizations hold). Note that depending on one’s philosophy of science, some fundamental laws of the universe may ultimately not be explainable.
Descriptive hypothesis	A proposed empirical generalization that describes (without explaining) a phenomenon or class of phenomena. If supported by evidence, then it may become a principle or rule or law. Although descriptive hypotheses can have predictive power insofar as their claims about *regularities* are well supported (and thus can be used to generate predictions), they are not predictions themselves.

guide rather than a presentation of universally agreed definitions. Still, scholars often find these concepts and distinctions useful for their heuristic and organizational value in discussions of theory (e.g., Gopnik & Wellman, 1994).

Key Elements of Theory

How Is Theory Linked to Reality?

Scientists and philosophers of science have grappled for decades with how theory and observations are linked (Godfrey Smith, 2003). All theories and hypotheses are necessarily linked to observations of the world, but there is disagreement about how theories relate to reality. There is also no formal theory specifying how theories ought to be evaluated or how theories can be securely arrived at from data.

Although operational definitions of concepts like *explanation* and *causation* that are at the heart of theory can be difficult to pin down, most scholars agree that theory comprises a key component of the scientific process, and it allows us to interpret, explain, and predict empirical phenomena. Furthermore, even though there is debate, there are some generally agreed principles for how to test and evaluate theories.

Some scientists and philosophers of science contend that Bayesian thinking may provide researchers with a formal theory of confirmation and evidence (see Earman, 1992 and Chapter 23 of this volume). In Bayesian thinking, two key ideas inform us of the probability that a hypothesis is true (Godfrey Smith, 2003). First, evidence (e) supports a hypothesis (h) only if e increases the probability of h. Second, probabilities are *updated* in accordance with Bayes's theorem – $P(h|e) = P(e|h)P(h)/[P(e|h)P(h)] + P(e|\text{not-}h)P(\text{not-}h)]$. To illustrate, imagine you are unsure whether reading this chapter will help you to become a better researcher. The hypothesis that the chapter will be helpful is h. Now imagine that you discover evidence e that informs you that the chapter is highly cited. Suppose now that before learning about the number of times the chapter has been cited, you estimated that the probability that this chapter would help you is 0.50. In other words, your initial estimate of the probability that this chapter would be helpful was 50%. Suppose that the probability of it being highly cited *given* that it is indeed helpful is 0.70 (in other words, imagine that 0.7 is the probability of finding e if h is true). Also suppose that the probability of the chapter being heavily cited if it is *not* helpful is only 0.20. Assuming that these prior probabilities are true, we can calculate the probability that the chapter will be helpful (i.e., the probability of h) given evidence that it is heavily cited. Using Bayes's theorem, we get $P(h|e) = (0.70)(0.50)/[(0.70)(0.50) + (0.20)(0.50)] = 0.77$. In other words, if we come across evidence that the chapter is highly cited, the probability of h goes up from 0.50 (our initial estimate) to 0.77. That is, Bayesian techniques can help us more accurately estimate the probability that a hypothesis is true as new evidence becomes available. Of course, this assumes we can accurately estimate the requisite prior probabilities for Bayes's theorem. That will sometimes be difficult, especially in the complex world of the social and behavioral sciences.

In null-hypothesis significance testing, we are always testing *P*(*e/h)* (technically, the probability of obtaining the *evidence* given *not-h* – the *null* hypothesis). This is in a sense backwards since what we *really* want to know is *P*(*h/e*). What we really want to know is: given the evidence we have obtained, what is the probability that our hypothesis is true? Bayes's theorem enables us to flip the question so that we are asking the question that we actually want answered. Although there is no universally accepted method for building theory, Bayes's theorem can render theories more tethered to reality by steering us toward the right questions and allowing us to more directly assess the probability that our hypotheses are correct.

How and When Should We Test Theory?

Most theoretically guided research in the social and behavioral sciences involves four steps: (a) generating causal hypotheses, (b) deriving predictions from those hypotheses, (c) empirically testing those predictions, and (d) interpreting the study results (Lewis et al., 2017). One way of conceptualizing this process is provided by Popper's (1959) hypothetico-deductive model. This consists of proposing a causal hypothesis and then testing predictions derived from the hypothesis with the goal of falsifying incorrect hypotheses (Popper, 1959). This "negative" rather than "positive" way of arriving at knowledge is considered a useful model for science, although some disagree about its utility and how accurately it describes the research activities of scientists (Borsboom et al., 2020; Godfrey Smith, 2003; Ketelaar & Ellis, 2000).

Many scholars have urged researchers not to feel pressured to generate and test causal hypotheses before they are ready (Barrett, 2020;Meehl, 1978; Rozin, 2001; Scheel et al., 2020). Some scientists caution us to first focus on (a) (re)conceptualizing the phenomena that we are interested in, (b) validating the constructs used to measure these phenomena, and (c) observing, cataloguing, and describing these phenomena before theorizing about them. In evolutionary biology, for example, decades of empirical research in taxonomy took place before formal phylogenetic theories were introduced (Nettle, 2021). Focusing on first improving measurements and amassing descriptions of phenomena and empirical generalities can lay the foundation for better theory and ensure we are not devising theories to explain inaccurate observations.

Theories Provide Explanations

One of the central roles of theory is to provide explanations. There is no universally agreed theory about the elements of a good explanation (Godfrey Smith, 2003). Explanations can be expected to take many forms because there is no single way to gauge explanatory goodness that works equally well in all scientific disciplines. Furthermore, a single phenomenon can often be explained in a number of different ways. Theory, however, can help us turn the sea of possible explanations into a smaller pool of more plausible ones. In addition to relevant information (i.e., signal), data contain information that is irrelevant to the phenomena of interest (i.e., noise). Theory helps us differentiate noise from signal and explain the phenomena of interest.

As the complexity of the phenomena we seek to explain increases, the pool of theories that can coherently explain the phenomena becomes progressively smaller (Dawkins, 1986). A complex phenomenon is one that involves many variables and causal connections, and it may require theory that is commensurately complex. The more causal propositions a theory posits and the more breadth we attempt to cover with our theory, the more we can explain (and the more that might go wrong). As the number of propositions increases, fewer other propositions can be added while maintaining internal coherence. As the theory's breadth increases, so do the possible ways in which evidence can counter the theory. Furthermore, as the complexity of the phenomena under study increases, so does the risk of overfitting (i.e., interpreting irrelevant noise as relevant signals and falling into the trap of "explaining" noise; see Gigerenzer, 2020). That is one reason why it is important to ensure that our theories can predict new findings (i.e., afford foresight) and not just explain data in hindsight.

Theories Incorporate Parallel Explanations and Multiple Levels of Analysis

Complex phenomena can often be explained or analyzed at multiple levels of analysis (Mayr, 1961; Tinbergen, 1963). For example, in the domain of biology and behavior, all phenomena can be analyzed and explained at four levels – also known as Tinbergen's four questions. They are: (1) survival value (i.e., adaptive function – how the trait contributes to survival and reproduction), (2) mechanism (i.e., causation – how a trait works mechanistically, including what triggers and regulates it), (3) development (i.e., ontogeny – how a trait develops over the lifespan), and (4) evolution (i.e., phylogeny – the evolutionary processes that gave rise to a trait).

The answers to Tinbergen's four questions offer four non-competing explanations of a trait, two of which are proximate and two of which are ultimate (see Nesse, 2013, p. 681, for a table that further organizes Tinbergen's four questions). From a theoretical perspective, this is important. First, recognizing that there are four *parallel* answers can correct misconceptions about competition between these different kinds of explanations. Second, recognizing complementary levels of analysis not only protects the researcher from contrived conflict but it can reveal gaps in theory (e.g., unexplored levels of analysis) and can lead to more complete explanations (Al-Shawaf, 2020). Third, the four questions are interrelated in ways that are useful and revealing when evaluating or proposing hypotheses or theories. For example, functional hypotheses yield specific predictions about proximate and mechanistic phenomena. An understanding of the latter can rule out certain functional hypotheses and point researchers toward others (Lewis et al., 2017).

The key point is that complex phenomena are often explicable at multiple levels of analysis. For a complete explanation of a mental phenomenon, we must address all four of Tinbergen's questions: how it evolved, why it evolved, how it works mechanistically, and how it developed across the organism's lifespan. These levels are typically non-competing. In other words, they are mutually compatible. When we ignore some levels, we fail to provide a comprehensive explanation of the phenomenon in question.

Characteristics of Good Theories

What makes a good theory? Theories vary on a number of characteristics, including simplicity, depth, breadth, and coherence. The best theories are often high in all four characteristics.

Simplicity or Parsimony

The principle of parsimony states that a theory should only posit entities that are *necessary* to do the explanatory work. One rule of thumb for building theories is to keep them as simple as possible. This does not mean that simple theories are more likely to be true than complex theories. A more complex theory is preferable to a simpler one if the simpler one is unable to explain the phenomena at hand; simplicity is most useful as a tiebreaker between theories that have the same explanatory and predictive power (Coelho et al., 2019). Simple theories are sometimes described as "elegant" and are said to benefit from "explanatory economy" (Tooby & Cosmides, 2015, p. 37).

Breadth

All else equal, a theory that can explain many different phenomena is preferable to a theory that can explain fewer phenomena. For example, a theory that can explain diverse behaviors across 1,000 species is more powerful than a theory that can do so across only 10 species. The more ground a theory covers, the greater the breadth of the theory.

A distinct kind of breadth involves the diversity of the *kinds* of evidence that support the theory (e.g., behavioral evidence, physiological evidence, cross-cultural evidence, and evidence from other species; Schmitt & Pilcher, 2004). For example, mating-related theories in humans were originally inspired by evidence from other species (Trivers, 1972). Subsequently, they were supported by evidence from humans across various cultures using psychological, physiological, and behavioral data (e.g., Buss, 1989). Convergent evidence from multiple sources enhances the likelihood that the theory is correct and raises our confidence in the veracity of the theory.

Depth

Depth here refers to explanatory depth. A theory is deeper if it provides *chains* of explanations rather than just a single explanation. Consider the following example: Why are men, on average, more violent than women? The answer is partly that, ancestrally, there was greater reproductive variance among men relative to women. In other words, men were more likely to be shut out of reproduction completely than women (a lower floor for reproductive success) but are also more capable of having a large number of offspring (a higher ceiling). As a consequence of this greater variance in reproductive success, aggression yielded greater reproductive payoffs for men than women. But why was there greater reproductive variance among men than among women in the first place? This is because of sex differences in the minimum parental investment in offspring. But why were there sex differences in the minimum parental investment in our species? This is partly due to sex differences in assurance of genetic parentage (maternity certainty and paternity uncertainty; Trivers, 1972). The point is that in this explanatory chain, we did not have to stop after explaining the initial phenomenon of interest; we were able to go deeper and *explain the explanation*. Explanatory depth can be increased by identifying the proximate causes of our initial phenomena of interest as well as the causes of those causes.

Coherence

People have used the term “coherence” to describe two characteristics of theory: (a) internal logical consistency and (b) accuracy – the latter of which refers to “coherence” with the external world. Empirical generalizations cannot be judged on internal coherence because they simply describe facts about the world. Theory, on the other hand, is evaluated on its internal coherence because it contains multiple propositions used to *explain* facts, and these propositions must be internally consistent. Internal coherence is thus achieved when an analysis demonstrates that the assumptions, propositions, and conclusions of a theory are logically consistent with one another. External coherence is achieved when an analysis demonstrates that the theory is consistent with other known principles or facts that are closely related to the theory in question. For example, the “crime and punishment model” (a theory positing that punishing crime is important for deterring crime; Becker, 1968) may be *internally* coherent, but it is not high in *external* coherence because it appears to be incompatible with the empirical findings criminologists have documented (De Courson & Nettle, 2021).

Coherence Stress-Tests

To increase the coherence of a theory that deals with complex phenomena, researchers can design “coherence stress-tests” to deliberately identify logical inconsistencies or incompatibilities within the theory or between the theory and data. This can be done in a number of ways, including attempts to disconfirm the theory and attempts to confirm, or be maximally charitable to, rival theories. This is an arduous process that some scholars find psychologically aversive because the process may involve a loss of prestige, among other costs, if one’s theories are disproven. Our human tendency to be more skeptical of viewpoints that contradict our beliefs can hinder the scientific enterprise (Greenwald et al., 1986). We need to counteract these tendencies by seeking and bolstering arguments that constructively criticize a theory, especially if it is one that we believe is true. It also helps to acknowledge facts that are apparently inconsistent with a favored theory. A perceived inconsistency between theory and data may sometimes lead us to abandon the theory or it may propel us to find ways to reconcile the two in a manner that improves or expands the theory. Darwin famously did this when he realized that his theory of natural selection could not explain the peacock’s lavish tail. Instead, it was his theory of sexual selection that eventually offered the explanation (Darwin, 1871).

Example of a Good Theory

The breadth, depth, and coherence of Darwin's theory of selection (natural and sexual), combined with the many sources of empirical evidence supporting it, make it the guiding paradigm of the life sciences. This theory explains known findings, predicts new ones, and integrates findings from a large variety of scientific fields (Al-Shawaf et al., 2018). The theory is also elegant and simple, as its main claim about evolution follows as a necessary conclusion given only three premises (genetic variation, inheritance, and differential reproduction). This is the closest thing that the social and behavioral sciences have to a universal scientific *law* (i.e., a regularity in nature that is universal; Dawkins, 1983).

Advantages of Theories

Good theories offer researchers several advantages. The more a theory exhibits the advantages discussed here, the more confidence we can have in its accuracy.

1. Explaining Findings That Are Otherwise Puzzling

One benefit of a good theory is that it can explain otherwise puzzling findings (Al-Shawaf, 2021). Atheoretical empirical work can *describe* puzzling phenomena but typically leaves these unexplained. To *explain* phenomena, especially in a psychologically satisfying way, we need theory (Gopnik & Wellman, 1994; Tooby & Cosmides, 1992). Good theories explain a phenomenon thoroughly, often across multiple levels of analysis.

2. Bridging Different Disciplines

Consilience, also known as conceptual or vertical integration, is the idea that findings across disciplines must not clash with one another. A consilient theory is consistent with the findings and theories of other disciplines. Contrary to what some believe, consilience does not entail reductionism. For example, the theory of natural selection is not reducible to theories in chemistry, and good theories in chemistry are not reducible to theories in physics, but they are all compatible with one another. Similarly, the social and behavioral sciences should be mutually compatible as well as compatible with the natural sciences and other disciplines related to the social sciences, including genetics, animal behavior, behavioral ecology, anthropology, and cognitive science (Tooby & Cosmides, 1992, 2015). This does not necessarily mean sociology is reducible to chemistry, but it does mean that the various sciences must not propose principles, hypotheses, and theories that *violate* those that are strongly supported in the other sciences.

The social and behavioral sciences often focus their studies on humans. Because humans are also biological creatures, the social and behavioral sciences can be thought of as nested within the larger umbrella of biology and the life sciences. As a result, we can borrow from successful biological theories such as the modern evolutionary synthesis – a paradigm that has proven extremely fruitful for the life sciences (Williams, 1966). As the geneticist Dobzhansky (1964, p. 449) famously remarked: "Nothing in biology makes sense except in the light of evolution." Although it may sound surprising to some social and behavioral scientists, proposing theories of human behavior or psychology that are incompatible with evolutionary biology is akin to proposing a chemical reaction that contradicts the laws of physics. Accordingly, social and behavioral scientists who want to ensure consilience and avoid obvious errors should make an effort not to run afoul of the principles and theories of evolutionary biology (Tooby & Cosmides, 1992, 2015).

Unfortunately, theoretical work in the social and behavioral sciences is often underdeveloped and may lack the breadth required to do the work of bridging different disciplines. Anthropologist Pascal Boyer once commented that "[t]he study of human behavior is encumbered by the ghosts of dead theories and paradigms" (Boyer, 2018, p. 28). These dead theories and paradigms do not have to encumber us, however, as they can narrow the search space by helping us rule out theories that failed to be supported by evidence or that failed to be consistent with established knowledge in other disciplines. The search for consilience is helpful in a similar way – it can narrow the search space by ruling out possibilities that are implausible given other disciplines' established findings and theories.

The social and behavioral sciences are replete with theories that have not been checked for compatibility with other areas within and beyond these fields, although there have been attempts to integrate related paradigms and theories (e.g., evolutionary and health psychology; Tybur et al., 2012). For example, researchers often limit themselves to the theories and empirical generalizations accepted in their specific departments, conferences, and journals (Gigerenzer, 2010). It is as if there were 10 separate investigations about one murder, but each investigative team was unconcerned with the hypotheses and findings of the other teams' investigations. If only they could consult each other and consolidate their theoretical analyses and findings, they would be more likely to uncover the answer. Fields that deal with a broad range of phenomena (i.e., human nature and culture) and are founded on theories and findings spanning several disciplines (e.g., cognitive science, anthropology, evolutionary biology, and psychology) exhibit greater consilience compared to fields that deal with a narrower range of phenomena or that engage with fewer theories from different disciplines. A shift toward greater interdisciplinarity can thus motivate the development of more accurate theory that explains a broader range of phenomena.

3. Predicting New Findings

Good theories lead to hypotheses that can make new predictions and lead us to new discoveries (Lewis et al., 2017; Muthukrishna & Henrich, 2019). In some cases, empirical generalizations can also help to accurately predict phenomena, and it is sometimes possible to use statistical relationships to predict phenomena without having a theory to explain them. Still, prediction is enhanced by good theory, and predicting findings in advance is a key means of assessing a theory's utility. Often, good theories and hypotheses will lead to predictions both related to what we expect to see in a given context or experiment as well as what we expect *not* to see. Finally, only theory (and not merely descriptive research) holds the promise of predicting new *kinds* of phenomena, as discussed next.

4. Pointing to Fruitful Questions

Good theory offers heuristic value – it can guide us in new and fruitful directions by hinting at the existence of previously unconsidered phenomena, even before looking at the data (Barrett, 2020; van Rooij & Baggio, 2021). It can also suggest new questions that we had not previously thought to ask. This advantage of theory – heuristic value – is not just about proposing a priori predictions. Instead, it is about asking new kinds of questions and starting new research areas that may have otherwise remained unexplored.

Ten Ways Theory Can Lead Us Astray

There is no question that the social and behavioral sciences should be grounded in good theory. However, it is also possible for theory to lead researchers astray, and we need to be aware of these pitfalls. In addition to the dangers posed by theory, we must also take our cognitive biases into account.

1. Seeking to Confirm Theory

Many psychological features of humans – such as confirmation bias or myside bias – hinder our search for truth and can affect how we conduct science. We selectively seek, remember, and attend to evidence that supports our beliefs (Lilienfeld, 2010; Loehle, 1987). We erroneously avoid theories that may contradict our ideological worldviews (e.g., von Hippel & Buss, 2017). We sometimes amend theory after the fact so that unexpected findings or counter-evidence can fall within its explanatory purview. Without knowing it, we may choose to observe or pay more attention to phenomena that confirm our hypothesis even when such findings are not especially helpful in testing our hypothesis (see, for example, the Wason selection task; Cosmides, 1989). Findings that are in line with predictions derived from our hypotheses support our hypotheses only *tentatively*. As a result, we need to be mindful of our tendency to seek confirmation of our hypothesis as well as our tendency to interpret data in ways that fit with our prior beliefs. The coherence stress-tests mentioned above can help to combat these tendencies.

2. Theory Influences How We Interpret Data

A theory's ability to guide us in interpreting data is one of the features that make theory useful. But if a theory is wrong, it can thwart our understanding of the data.

Because we may be motivated to interpret data in ways that confirm our theory, the risk of misinterpreting data may be considered a manifestation of the problem of seeking to confirm theory, discussed above. Even descriptive findings, similar to empirical generalities, are subject to our confirmation bias-infused interpretations. However, having a specific theory in mind before one begins increases this risk. As fictional detective Sherlock Holmes put it, "[i]t is a capital mistake to theorize before you have all the evidence. It biases the judgment" (Doyle, 1887/1995, p. 23).

3. Theory Influences How We Measure and What We Observe

When theory influences how we make observations, or what we choose to observe, these observations are said to be theory-laden. This means that our findings may be biased by our previously held theoretical beliefs or folk intuitions (Lilienfeld, 2010). Whenever feasible, it is important to be transparent about how theory may have influenced our measures, constructs, and interpretations of our findings, although this may sometimes be unconscious (Barrett, 2020). Theory can also influence observation in the sense that our theories tell us where to look and what is worth observing in the first place (e.g., Barrett, 2020). To the extent that we are burdened with an invalid theory, we may be wasting time by observing or measuring the wrong things.

4. Poorly Defined Theoretical Constructs

It is necessary to define our theoretical concepts with as much precision as possible (Gerring, 1999). What are the necessary and sufficient attributes of the phenomena under study, if any? How differentiated are the constructs that capture these attributes from similar concepts? For example, theories that claim to differentiate grit from conscientiousness may need to be revised given meta-analytic evidence that these two concepts are highly interrelated (Credé et al., 2017). If the concepts and variables included in our theoretical work are not well specified or operationally defined, we will be unable to gauge whether our measures are behaving as expected. The concept "social group," for instance, is ubiquitous and yet difficult to operationalize (Pietraszewski, 2021). This is problematic because it can give us more leeway when interpreting findings and may leave us more vulnerable to the problem of accommodating unanticipated findings in our theory.

5. Theorizing Too Soon

Are we proposing causal hypotheses and theories too soon? Journals that encourage theory are no doubt useful for the social and behavioral sciences. At the same time, the review process for some journals in these fields may be pushing us to theorize too soon (and possibly unduly criticize manuscripts that do not offer much in the way of theory; Biswas-Diener & Kashdan, 2021). It may be helpful to keep in mind the risk that our eagerness to theorize about phenomena sometimes exceeds our ability to realistically do so in a rigorous way (Barrett, 2020).

In some cases, we may be theorizing too early in the sense that we are attempting to explain phenomena that are not yet properly described. In such situations, accurate descriptive empirical work can be a crucial foundation before theoretical explanations are attempted (Barrett, 2020; Rozin, 2001; van Rooij & Baggio, 2021). We note that there are many books, articles, and courses that teach us how to conduct empirical research in the social and behavioral sciences, but the same is much less true for theoretical research and theory construction (Gray, 2017; despite some exceptions, Fried, 2020). A useful exception is Borsboom et al.'s (2020) course about building theory with practical suggestions for developing an interdisciplinary understanding of the phenomena of interest.

6. Theorizing Too Late

Hypothesizing after the results are known (HARKing) is the process of revising a hypothesis after we have looked at the data so that the hypothesis can better account for the data – especially data that do not fit well with the original hypothesis (Kerr, 1998). To be charitable, HARKing may be an indication of non-fraudulent scenarios, including: (1) the unpredicted findings *could* have been predicted via the original hypothesis but the researcher simply forgot to derive the prediction that would have forecasted the unanticipated data or (2) the hypothesis really does need to be amended to incorporate the unanticipated findings because these could be explained by an amended causal hypothesis (but this must be done transparently). However, HARKing can also be an indication that (3) our hypothesis can too easily accommodate all kinds of data because it is underspecified or unfalsifiable or (4) we are engaging in the epistemologically and ethically suspect behavior of pretending we predicted something in advance when we did not. In the social and behavioral sciences, theories are often formulated in such an unspecified and loose way that it is nearly impossible for any finding to disconfirm them (Meehl, 1978). Unspecified theories are more amenable to HARKing-type revisions that sometimes take the form of positing the existence of moderator variables that would make the hypothesis more compatible with the data.

7. Not Even Theory

One way to avoid being led astray by theory is to learn about the common but loose surrogates that masquerade as theory in the social and behavioral sciences. As discussed by Gigerenzer in his short essay on the subject, these surrogates include labels (e.g., "cultural," "learned," and "evolved"), false dichotomies (e.g., learned versus evolved), and underdeveloped theoretical concepts and connections (Gigerenzer, 2009). A complement to HARKing is CITEing (calling it theory for effect), which is when we call something a theory even though we are referring to empirical generalities (Nettle, 2021). It is often better to delay or avoid proposing a theory than it is to propose one that is vague and underspecified. For instance, a theory needs to specify the domains to which it applies as well as those to which it does not (Gigerenzer, 2020).

8. Vagueness, Imprecision, and the Utility of Formal Mathematical Models

"Formalized" theory is a theory that is quantified and uses mathematics to increase precision (Guest & Martin, 2021). Mathematical models have some advantages over verbal models: (1) they are often explicit about the assumptions that they make, (2) they are precise about the constructs that they use, and (3) they may make it easier to derive predictions from the hypothesis (Guest & Martin, 2021; Smaldino, 2020). As a result, theories that are formalized with mathematical models are sometimes more transparent about the assumptions and relationships included in the model.

There are many benefits to becoming familiar with mathematical models. Putting on a "modeler's hat" can improve our ability to think clearly (Tiokhin, 2021). Epstein (2008) lists numerous reasons to build mathematical models, including developing causal explanations, suggesting analogies, demonstrating trade-offs, and revealing the apparently simple to be complex (and vice versa). In the absence of mathematical modeling, using verbal qualifiers (i.e., phrases expressing the degree of confidence one has in an assumption or verbally delineating the boundary conditions of the phenomenon) can also serve to promote better theory specification and transparency.

9. Theory Can Send Us Down the Wrong Paths and Waste Our Time

Our eagerness to theorize, combined with the way theories guide our thinking, may lead us to ask the wrong questions and waste time pursuing unfruitful research. This problem is exacerbated when we are overly confident in our theory. In some situations, a bottom-up, observation-driven approach may be preferable to a top-down approach in which an invalid theory dictates where we should look and which research questions we should ask. Additionally, the hypothetico-deductive model's popularity may lead us to focus too much on (dis)confirming causal hypotheses at the expense of other key components of the scientific process that often need to precede or complement the testing of causal hypotheses (Borsboom et al., 2020). As discussed earlier, amassing descriptions of phenomena and identifying empirical generalities can be a useful starting point and stepping stone for theoretical work.

10. Missing Out on Phenomena

Top-down research begins with theory, whereas bottom-up research begins with observation. A top-down account of a phenomenon has the strength of generating a priori predictions. The bottom-up approach can sometimes be prone to post hoc explanations if not executed properly. Still, bottom-up approaches are an important source of knowledge about the world, and the risk of post hoc explanation can be avoided if we derive (and test) new predictions from the hypothesis we just put forth to explain our bottom-up observations (Al-Shawaf, 2020; Al-Shawaf et al, 2018). Briefly put, the risks of bottom-up research can be mitigated, and theory-driven top-down research has its costs, too. That is, if our research is derived top-down from theory, and our theory doesn't point toward a particular phenomenon, we may miss certain phenomena.

Ways to Develop Theory

Integrating What We Already Know

Integrating Theoretical and Empirical Work

Connecting theories is one way to increase our ability to explain otherwise puzzling findings. A theory integration program can take the form of two simple steps that build on each other (Gigerenzer, 2017). The first step involves the integration of empirical findings that are each explained by their own theories. The second step involves the integration of these otherwise disconnected theories. Gigerenzer (2008) has suggested that integration can take the form of collating two existing theories, and he cites as an example the productive merger between the ACT-R cognitive architecture program and the Adaptive Toolbox program – a merger that led to a counterintuitive "less-is-more" discovery that simpler heuristics can yield better results than more computationally intensive procedures (Schooler & Hertwig, 2005).

Integrating Theory with Methodology

Integrating theory with the methods that we use can help us to develop and improve theory. Reliable methods and good theories are synergistic in the sense that (1) theories can suggest new methods and (2) new methods allow access to previously unreachable findings that can inspire new theories or refine existing ones (Gray, 2017). Reliable methods and rigorous theory can inspire improvements to one another.

Thinking About Psychology and Behavior Across Three Computational Stages

To illustrate top-down theorizing (i.e., generating a priori causal hypotheses), consider evolutionary psychology, which draws from both evolutionary theory and the computational sciences (Tooby & Cosmides, 2015). For example, one can approach psychology and behavior with a three-step model borrowed from the cognitive sciences. These three steps are the "inputs" stage, the "processes" stage, and the "outputs" stage. The "inputs" stage is when we specify the stimuli that a psychological mechanism is predicted to be sensitive to (i.e., the inputs that the mental mechanism is expected to process). In this first step, it is also useful to specify the inputs that the trait is predicted *not* to be sensitive to (i.e., the inputs predicted to be irrelevant; see Lewis et al., 2017 for a discussion of such "negative" predictions). The "processes" stage involves identifying the algorithms and decision rules by which the psychological mechanism processes the relevant inputs. The “outputs” stage – the stage perhaps most familiar to social and behavioral scientists – involves specifying the behavioral, cognitive, and physiological characteristics that the mental trait produces as outcomes. This last stage can be thought of as the outcome of the first two stages.

This model can help us to avoid gaps in our understanding of psychological and behavioral phenomena. These gaps often reside in the “processes” stage that was often ignored by the behaviorists (Norris & Cutler, 2021), who focused solely on the stimulus stage (roughly, the inputs) and the response stage (roughly, the outputs). In sum, this model can be a useful reminder not to elide processing stage between inputs and outputs.

Other Conceptual Tools

There are other conceptual tools for theory building at our disposal. A tool called “condition seeking” describes the act of identifying the necessary and sufficient conditions of a phenomenon. It involves asking questions such as “is the phenomenon domain-general or domain-specific?” and “have we exhausted the conditions under which this phenomenon emerges?” (Greenwald et al., 1986). Another tool at our disposal involves “reverse-engineering,” which is useful for generating hypotheses about why certain psychological capacities exist or why they work the way that they do (Tooby & Cosmides, 1992). Consider, for example, friendship jealousy. A third key conceptual tool is called “evolutionary task analysis,” which (a) begins with an “adaptive problem” humans have recurrently faced during their evolution, (b) asks what kind of psychological mechanism could possibly solve such a problem, and then (c) posits hypotheses about how this psychological mechanism might work (e.g., see Al-Shawaf et al., 2016; Lewis et al., 2017). For a longer discussion of useful conceptual tools, see Kenrick’s (2020) table listing six heuristics for generating hypotheses along with examples and applications for each heuristic.

One Last Red Flag: Too Much Explaining and Too Little Predicting

Lakatos (1976) argued that a research area can be said to be progressing when its theoretical growth *anticipates* its empirical growth. That is, as long as it demonstrates predictive power by helping us to generate novel empirical findings. By contrast, it is “degenerative” or stagnant if its theoretical growth lags behind its empirical growth. As a result, too much explaining and too little predicting is the kind of lag that scientists may regard as a red flag. To check the discrepancy between a theory’s explanatory and predictive power, we need to first examine a research field with an eye to the number of novel findings predicted by the theory. To do this properly, we may need to control for factors such as the number of researchers who use the theory, the resources they have at their disposal, and how long the theory or research field has been active (Miller, 2000). The key point is that we need to be aware of how much post hoc explaining is occurring relative to a priori theorizing.

At present, theories in the social and behavioral sciences often do too much explaining and too little predicting (Yarkoni & Westfall, 2017). Theory is often amended to explain findings and empirical generalities that were not predicted a priori. Finding counter-evidence to a theory sometimes leads researchers to (a) reinterpret the counter-evidence as consistent with the theory (also referred to as conceptual stretching; Scheel et al., 2020) or (b) treat the counter-evidence as irrelevant noise (Lakatos, 1976). Such a posteriori revising of theory to accommodate findings risks making our theories less coherent. Furthermore, a theory that can explain everything may not be explaining anything. As a result, post hoc explanations must be regarded with caution (Ellis & Ketelaar, 2000), and new predictions must be derived (and then tested) from the recently posited post hoc explanations as a key "check" or safeguard (Al-Shawaf et al., 2018).

Despite these dangers, a posteriori revising can sometime be important in building good theories. As discussed earlier, it is sometimes better to revise a theory after finding counter-evidence rather than getting rid of the theory altogether – the latter may be going too far (see the subsection above on coherence). This is one of the central tensions of science – it is important to revise one's theory in accordance with counter-evidence, but it is also important not to have a theory that is infinitely malleable and stretchy, capable of accommodating anything (and therefore explaining nothing). These tensions and balances are often a key part of science.

Conclusions and Summary of Theory-Related Recommendations

To conclude, good theory helps generate hypotheses as well as narrow them down, and it has great utility in helping us more efficiently interpret, explain, and predict phenomena in the world (Muthukrishna & Henrich, 2019). Theory is thus extremely useful and can spark progress in the currently disunited and often atheoretical social and behavioral sciences. At the same time, theorizing contains risks because theory can bias what we see, where we choose to look, and how we interpret our results. Of course, if our theory is reliably explanatory and predictive, then this effect will be positive – it will help us to *more correctly* interpret what we see, suggest useful new directions for research, and lead to plausible new predictions. Seen in this way, using theory is a high-risk, high-reward game for scientists who are trying to improve their empirical research and their understanding of the world.

To improve research in the complex realm of the social and behavioral sciences, it may be useful for us to remember the following theory-related recommendations:

- Delay theoretical work until we have better concepts, methods, and empirical descriptions (see "How and When Should We Test Theory?").
- Ensure that our theories predict new findings, not just explain known ones (see "Predicting New Findings and Pointing to Fruitful Questions").
- Specify what our theory predicts will (and will *not*) occur and consider computational three-stage models of psychology and behavior for more complete theories

(see "Thinking About Psychology and Behavior Across Three Computational Stages").

- Consider complementary levels of analysis for more complete theories (see "Theories Incorporate Parallel Explanations and Multiple Levels of Analysis").
- Remember that parsimony and simplicity are important, but more complex theories may be needed if simpler theories are unable to explain the phenomena of interest (see "Simplicity or Parsimony").
- Strive to improve theories' breadth, depth, and coherence (see "Breadth," "Depth," and "Coherence").
- Diversify sources of evidence (see "Breadth").
- Integrate theoretical work across disciplines to ensure consilience (see "Bridging Different Disciplines").
- Formalize theoretical structures with mathematical modeling or verbal qualifiers for more precision and transparency (see "Vagueness, Imprecision, and the Utility of Formal Mathematical Models").

Engaging with predictively and explanatorily powerful theories will put the social and behavioral sciences on firmer footing, but it is not a magic bullet. Consider the possibility of needing to scrutinize two detectives' stories to determine whose "theory" should be prioritized in a murder investigation on a tight budget. It may not be enough to only scrutinize the detectives' methods and tools. Scrutinizing the plausibility of their theories or hypotheses on the basis of the criteria discussed in this chapter may also be helpful, though perhaps not enough. We may additionally need to consider a number of miscellaneous factors such as the detectives' (a) confidence levels in their claims, (b) personality traits that lead them to be overconfident or underconfident in their judgments, (c) past efficiency in solving similar problems, (d) intellectual honesty, and (e) degree of rigor, clarity, and nuance when making claims. These kinds of factors are studied by philosophers of science and sociologists of science to better understand how our procedures, psychologies, incentive structures, and values may be helping or hindering the scientific enterprise (e.g., Merton, 1973).

Science is one of the humankind's most powerful inventions (Borsboom et al., 2020), and theory is a key part of science. Theory can not only drive empirical work, but it also has the unique ability to help researchers interpret and explain phenomena, predict the existence of novel phenomena, and link bodies of knowledge. Theory offers heuristic value. It can steer us in directions that we otherwise would not have traveled. However, theory can also steer us away from the truth, given its ability to affect how we measure and what we observe, bias our interpretations, and cause us to waste time and resources by leading us down incorrect paths. Additional dangers stem from "surrogate theories," seeking to confirm theory, and theories that are so loosely specified that they can accommodate unanticipated findings. Despite these potential pitfalls, strong theories hold immense promise for the social and behavioral sciences. To build and assemble robust theories and bodies of knowledge in the social and behavioral sciences, cross-pollination of different theoretical and empirical research programs is key. This kind of scientific progress holds great potential for achieving two grand goals: increasing our understanding of the world and reducing the suffering of humans and other sentient beings (Kenrick, 2020; Gainsburg et al., 2021).

Acknowledgments

We thank the editors for inviting us to write this chapter on theory and for their constructive comments on an earlier version of this chapter.

References

Al-Shawaf, L. (2020). Evolutionary psychology: Predictively powerful or riddled with just-so stories? *Areo Magazine*, October 20. Available at: https://areomagazine.com/2020/10/20/evolutionary-psychology-predictively-powerful-or-riddled-with-just-so-stories/.

Al-Shawaf, L. (2021). Evolution explains puzzling aspects of the human mind: Evolution helps explain anxiety, the hedonic treadmill, and other puzzles. *Psychology Today*, July 11.

Al-Shawaf, L., Conroy-Beam, D., Asao, K., & Buss, D. M. (2016). Human emotions: An evolutionary psychological perspective. *Emotion Review*, *8*(2), 173–186. https://doi.org/10.1177/1754073914565518

Al-Shawaf, L., Zreik, K. A., & Buss, D. M. (2018). Thirteen misunderstandings about natural selection. In T. K. Shackelford & V. A. Weekes-Shackelford (eds.), *Encyclopedia of Evolutionary Psychological Science* (pp. 1–14). Springer.

Barrett, H. C. (2020). Deciding what to observe: Thoughts for a post-WEIRD generation. *Evolution and Human Behavior*, *41*(5), 445–453.

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy*, *76*(2), 169–217. https://doi.org/10.1007/978-1-349-62853-7_2

Biswas-Diener, R. & Kashdan, T. B. (2021). Three lessons from Ed Diener. *International Journal of Wellbeing*, *11*(2), 73–76. https://doi.org/10.5502/ijw.v11i2.1705

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Borsboom, D., van der Maas, H. L. J., Dalege, J., Kievit, R., & Haig, B. (2020). Theory construction methodology: A practical framework for theory formation in psychology. *Perspectives on Psychological Science*, *16*(4), 756–766. https://doi.org/10.1177/1745691620969647

Boyer, P. (2018). *Minds Make Societies: How Cognition Explains the World Humans Create*. Yale University Press. https://doi.org/10.12987/9780300235173

Burns, M. K. (2011). School psychology research: Combining ecological theory and prevention science. *School Psychology Review*, *40*(1), 132–139. https://doi.org/10.1080/02796015.2011.12087732

Buss, D. M. (1989). Sex differences in human mate preferences: Evolutionary hypotheses tested in 37 cultures. *Behavioral and Brain Sciences*, *12*(1), 1–14. https://doi.org/10.1017/S0140525X00023992

Coelho, M. T. P., Diniz-Filho, J. A., & Rangel, T. F. (2019). A parsimonious view of the parsimony principle in ecology and evolution. *Ecography*, *42*(5), 968–976. https://doi.org/10.1111/ecog.04228

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, *31*(3), 187–276. https://doi.org/10.1016/0010-0277(89)90023-1

Credé, M., Tynan, M. C., & Harms, P. D. (2017). Much ado about grit: A meta-analytic synthesis of the grit literature. *Journal of Personality and Social Psychology*, *113* (3), 492–511. https://doi.org/10.1037/pspp0000102

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Darwin, C. (1871). *The Descent of Man, and Selection in Relation to Sex*. John Murray.

Dawkins, R. (1983). Universal Darwinism. In M. A. Bedau & C. E. Cleland (eds.), *The Nature of Life: Classical and Contemporary Perspectives from Philosophy and Science* (pp. 403–425). Cambridge University Press.

Dawkins, R. (1986) *The Blind Watchmaker*. Norton.

De Courson, B. & Nettle, D. (2021). Why do inequality and deprivation produce high crime and low trust? *Scientific Reports*, *11*(1), 1–27. https://doi.org/10.31234/osf.io/p2aed

Dobzhansky, T. (1964). Biology, molecular and organismic. *American Zoologist*, *4*(4), 443–452. https://doi.org/10.1093/icb/4.4.443

Doyle, A. C. (1887; electronic reprint 1995). *A Study in Scarlet*. J. H. Sears & Co.

Earman, J. (1992). *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. MIT Press.

Ellis, B. J. & Ketelaar, T. (2000). On the natural selection of alternative models: Evaluation of explanations in evolutionary psychology. *Psychological Inquiry*, *11*(1), 56–68. https://doi.org/10.1207/S15327965PLI1101_03

Epstein, J. M. (2008). Why model? *Journal of Artificial Societies and Social Simulation*, *11* (4), Article 12. http://jasss.soc.surrey.ac.uk/11/4/12.html

Fried, E. I. (2020). Theories and models: What they are, what they are for, and what they are about. *Psychological Inquiry*, *31*(4), 336–344. https://doi.org/10.1080/1047840X.2020.1854011

Gainsburg, I., Pauer, S., Abboub, N., Aloyo, E. T., Mourrat, J. C., & Cristia, A. (2022). How effective altruism can help psychologists maximize their impact. *Perspectives on Psychological Science*. Advance online publication. https://doi.org/10.1177/17456916221079596

Gerring, J. (1999). What makes a concept good? A criterial framework for understanding concept formation in the social sciences. *Polity*, *31*(3), 357–393.

Gigerenzer, G. (2008). Why heuristics work. *Perspectives on Psychological Science*, *3*(1), 20–29. https://doi.org/10.1111/j.1745-6916.2008.00058.x

Gigerenzer, G. (2009). Surrogates for theory. *Association for Psychological Science Observer*, *22*, 21–23.

Gigerenzer, G. (2010). Personal reflections on theory and psychology. *Theory & Psychology* *20*(6), 733–743. https://doi.org/10.1177/0959354310378184

Gigerenzer, G. (2017). A theory integration program. *Decision*, *4*, 133–145. https://doi.org/10.1037/dec0000082

Gigerenzer, G. (2020). How to explain behavior? *Topics in Cognitive Science*, *12*(4), 1363–1381. https://doi.org/10.1111/tops.12480

Goetz, A. T. & Shackelford, T. K. (2006). Modern application of evolutionary theory to psychology: Key concepts and clarifications. *American Journal of Psychology*, *119*, 567–584. https://doi.org/10.2307/20445364

Godfrey Smith, P. (2003). *Theory and Reality: An Introduction to the Philosophy of Science*. University of Chicago Press.

Gopnik, A. & Wellman, H. M. (1994). The theory theory. In L. A. Hirschfeld & S. Gelman (eds.), *Mapping the Mind: Domain Specificity in Cognition and Culture* (pp. 257–293). Cambridge University Press. https://doi.org/10.1017/CBO9780511752902.011

Gray, K. (2017). How to map theory: Reliable methods are fruitless without rigorous theory. *Perspectives on Psychological Science*, *12*(5), 731–741.

Greenwald, A. G., Pratkanis, A. R., Leippe, M. R., & Baumgardner, M. H. (1986). Under what conditions does theory obstruct research progress? *Psychological Review*, *93*(2), 216–229.

Guest, O. & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*(4), 789–802. https://doi.org/10.1177/1745691620970585

Kenrick, D. T. (2020). Discovering the next big question in evolutionary psychology: A few guidelines. *Evolutionary Behavioral Sciences*, *14*(4), 347–354.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217.

Ketelaar, T. & Ellis, B. J. (2000). Are evolutionary explanations unfalsifiable? Evolutionary psychology and the Lakatosian philosophy of science. *Psychological Inquiry*, *11*(1), 1–21.

Lakatos, I. (1976). Falsification and the methodology of scientific research programmes. In S. G. Harding (ed.), *Can Theories Be Refuted?* (pp. 205–259). Springer.

Lewis, D. M. G., Al-Shawaf, L., Conroy-Beam, D., Asao, K., & Buss, D. M. (2017). Evolutionary psychology: A how-to guide. *American Psychologist*, *72*(4), 353–373.

Lilienfeld, S. O. (2010). Can psychology become a science? *Personality and Individual Differences*, *49*(4), 281–288.

Loehle, C. (1987). Hypothesis testing in ecology: Psychological aspects and the importance of theory maturation. *The Quarterly Review of Biology*, *62*(4), 397–409.

Mayr, E. (1961). Cause and effect in biology. *Science*, *134*(3489), 1501–1506.

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.

Merton, R. K. (1973). *The Sociology of Science: Theoretical and Empirical Investigations*. University of Chicago Press.

Miller, G. (2000). How to keep our metatheories adaptive: Beyond Cosmides, Tooby, and Lakatos. *Psychological Inquiry*, *11*(1), 42–46.

Muthukrishna, M. & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*, 221–229.

Nesse, R. M. (2013). Tinbergen's four questions, organized: A response to Bateson and Laland. *Trends in Ecology & Evolution*, *28*(12), 681–682.

Nettle, D. (2021). Theories and models are not the only fruit. *Blog post*. Available at: https://leonidtiokhin.medium.com/theories-and-models-are-not-the-only-fruit-a05c7cf188f6.

Norris, D. & Cutler, A. (2021). More why, less how: What we need from models of cognition. *Cognition*, *213*, 104688. 10.1016/j.cognition.2021.104688

Pietraszewski, D. (2021). Towards a computational theory of social groups: A finite set of cognitive primitives for representing any and all social groups in the context of conflict. *Behavioral and Brain Sciences*, April 27, 1–62. https://doi.org/10.1017/S0140525X21000583

Popper, K. R. (1959). *The Logic of Scientific Discovery*. Basic Books.

Poincaré, H. (1905). *Science and Hypothesis*. Science Press.

Rozin, P. (2001). Social psychology and science: Some lessons from Solomon Asch. *Personality and Social Psychology Review*, *5*(1), 2–14.

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2020). Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science*, *16*(4): 744–755. https://doi.org/10.1177%2F1745691620966795

Schmitt, D. P. & Pilcher, J. J. (2004). Evaluating evidence of psychological adaptation: How do we know one when we see one? *Psychological Science*, *15*(10), 643–649.

Schooler, L. J. & Hertwig, R. (2005). How forgetting aids heuristic inference. *Psychological Review*, *112*(3), 610–628. https://doi.org/10.1037/0033-295X.112.3.610

Smaldino, P. E. (2020). How to translate a verbal theory into a formal model. *Social Psychology*, *51*(4), 207–218. https://doi.org/10.1027/1864-9335/a000425

Symons, D. (1992). On the use and misuse of Darwinism in the study of human behavior. In J. H. Barkow, L. Cosmides & J. Tooby (eds.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture* (pp. 137–159). Oxford University Press.

Tinbergen, N. (1963). On aims and methods of ethology. *Zeitschrift für Tierpsychologie*, *20*(4), 410–433.

Tiokhin, L. (2021). Models are for transparency. *Blog post*. Availabale at: https://doi.org/10.13140/RG.2.2.11024.53767.

Tooby, J. & Cosmides, L. (1992). The psychological foundations of culture. In Barkow, J.H., Cosmides, C., & Tooby, J. (eds.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture* (pp. 19–136). Oxford University Press.

Tooby, J. & Cosmides, L. (2015). The theoretical foundations of evolutionary psychology. In Buss, D. M. (ed.), *The Handbook of Evolutionary Psychology, Second Edition. Volume 1: Foundations*. (pp. 3–87). John Wiley & Sons.

Trivers, R. L. (1972). Parental investment and sexual selection. In B. Campbell (ed.), *Sexual Selection and the Descent of Man: 1871–1971* (pp. 136–179). Aldine.

Tybur, J. M., Bryan, A. D., & Hooper, A. E. C. (2012). An evolutionary perspective on health psychology: New approaches and applications. *Evolutionary Psychology*, *10*(5), 855–867.

van Rooij, I. & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science*, *16*(4), 682–697. https://doi.org/10.1177/1745691620970604

von Hippel, W. & Buss, D. M. (2017). Do ideological driven scientific agendas impede understanding and acceptance of evolutionary principles in social psychology? In J. T. Crawford & L. Jussim (eds.), *The Politics of Social Psychology* (pp. 7–25). Psychology Press.

Williams, G. C. (1966). *Adaptation and Natural Selection*. Princeton University Press.

Yarkoni, T. & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

2 Research Ethics for the Social and Behavioral Sciences

Ignacio Ferrero and Javier Pinto

Abstract

This chapter explores the nature of the work that researchers in the social and behavioral sciences do through a discussion of the ethical principles that ought to guide their work. Since academic researchers have different perceptions and attitudes regarding what constitutes (un)ethical research, we offer an overview of what is considered best practices in social and behavioral science research. This work focuses primarily on the ethical issues related to the design, development, implementation, and publication of research projects. It concludes with a guide for assisting research teams and research ethics committees in assessing the honesty, authenticity, and accountability of their research programs.

Keywords: Research Ethics, Integrity, Honesty, Accountability, Authenticity, Codes of research, Care

Ethics for Professional Research

Social and behavioral science research aims to understand human behavior in society and to produce useful knowledge. However, such knowledge can only have a positive impact on the well-being of society if it is acquired in accordance with the norms of scientific inquiry, assuring the reliability and validity of the indicators and outcomes and respecting the dignity of individuals, groups, and communities (Social Research Association, 2003). As such, the social and behavioral sciences can be a means to a better world if the knowledge that it produces is informed by ethical and responsible research (RRBM, 2017).

The ethics of research practices cannot be reduced to a checklist of standards and specific norms. In addition to complying with bureaucratic and administrative procedures, a research program is concerned with the integration of the basic goal of academic practice and knowledge, with the essential principles associated with human rights, and the common good of society. For instance, the Canadian Panel of Research Ethics has based their ethical standards on the respect for personal autonomy; the concern for personal welfare – the quality of experiencing life in all its aspects (e.g., physical, mental, and spiritual health); and justice (i.e., treating people fairly and equitably with equal respect and concern).

Historically, the concern for research practices was raised due to the research activities of William Beaumont during the 1830s and the leprosy studies conducted by Doctor Hansen during the end of the nineteenth century (Lock, 1995). In recent history, modern research ethics and the beginning of institutional review boards originated from the lived experiences in Nazi concentration camps that were revealed during the Nuremberg trials. This led to the Geneva Declaration in 1947 and the Helsinki Declaration in 1964 (Ruyter, 2003). However, even though different associations and academic organizations promote better ethical practices, fraud and malpractice in professional research are often written about in the press. From as early as the 1970s, the social and behavioral sciences seem to have suffered a wave of incidents of scientific misconduct, and scholars have documented the prevalence of questionable research practices (RRBM, 2017). A number of experiments on human subjects in the United States during the 1960s and 1970s also sparked a public outcry. Not only were they declared unethical and illegal since they were performed without the knowledge or informed consent of the test subjects, but they also set in motion a national debate that would eventually lead to stricter controls governing medical and social/behavioral human research. The most infamous of these experiments, the Tuskegee syphilis experiment, the MK-Ultra project, the Monster Study, the Stanford Prison experiment, and the Milgram experiment, demonstrated the extent to which human dignity and rights could be violated in the name of research (Brandt, 1978; Grimes et al., 2009; Zimbardo et al., 1971).

In 1997, Dotterweich and Garrison administered a survey to a sample of professors at institutions accredited by the Association to Advance Collegiate Schools of Business (AACSB) to identify those actions that they felt were unethical and to gauge the state of research ethics among business academics. The survey was centered on 11 substantive issues concerning business research ethics. More than 95 percent of respondents condemned 5 out of the 11 activities studied, including falsifying data, violating confidentiality of a client, ignoring contrary data, plagiarism, and failing to give credit to co-authors (Dotterweich & Garrison, 1997). Later, in 2011, the National Academy of Sciences showed a 10-fold increase in retractions and related misconduct over the past decades (Wible, 2016). In all, this brings into question both the research practices and beliefs of social and behavioral scientists and the work they produce.

These experiences highlight the fact that the science community has attracted the attention of public authorities who brought into question the reputability of research. For this reason, federal agencies in the United States now require systematic research ethics training programs as a mandatory part of the grant submission process. Similarly, the European Union's Horizon 2025 program encourages applicants to embed research ethics within their proposals (ERC, 2021).

Moreover, the concern for good research practice has led to the emergence of professional bodies whose role is to specify the rules and regulations that govern best practices, including codes of conduct and ethical committees in research institutions. Indeed, the proliferation of research codes of practice in use at academic institutions and research centers worldwide is further proof of the commitment to create a culture of best practices in social and behavioral research. Some examples of the former are the Singapore Statement on Research Integrity, the European Code of Conduct for Research Integrity, the World Medical Association's Declaration of Helsinki, the National Institutes of Health (NIH), and the National Science Foundation (NSF). Other influential research ethics policies, such as the Code of Ethics of the American Psychological Association, the Ethical Principles of Psychologists and Code of Conduct, the Statements on Ethics and Professional Responsibility (American Anthropological Association), and the Statement on Professional Ethics (American Association of University Professors), are also important to this effort.

In March 2014, after a series of symposia, conferences, and meetings, the Council of the Academy of Social Sciences formally adopted five guiding ethical principles for social science research (Dingwall et al., 2017). These are:

(i) Social science is fundamental to a democratic society and should be inclusive of different interests, values, funders, methods, and perspectives.
(ii) All social sciences should respect the privacy, autonomy, diversity, values, and dignity of individuals, groups, and communities.
(iii) All social sciences should be conducted with integrity throughout, employing the most appropriate methods for the research purpose.
(iv) All social scientists should act with regard to their social responsibilities in conducting and disseminating their research.
(v) All social science should aim to maximize benefits and minimize harm.

What these codes and principles have in common is that they adopt and promulgate preventative strategies that aim to produce more reliable and actionable knowledge for better policies and practices. In so doing, they contribute to a sense of professional practice integrally linked to compliance with rules and regulations and not just the authentic fulfilment of a professional role per se (OECD, 2007).

Nevertheless, conducting ethical research cannot only be about following specific rules. It must also incorporate a set of moral principles (British Psychological Society, 2021). This moral reasoning is essential and forms the basis of professional practice and meaningful work (Kanungo, 1992). In social and behavioral science research, moral reasoning must reinforce ethical principles to sustain the ethical research ecosystem (Drenth, 2012). Thus, an ethical culture of social and behavioral science research must encompass: (i) general principles of professional integrity and (ii) the principles applying to each of the constituencies that comprise the social ecosystem in which scientists work (Bell & Bryman, 2007).

In this chapter, we address some general principles of professional integrity in research; namely, honesty, objectivity, accountability, authenticity, compliance, and care. Then, we describe ethical principles that can guide institutions and other constituencies in their professional research practices. We discuss the role of research institutions as responsible employers, the responsibility of the scientific community to safeguard good practices, the role of public and private institutions to manage research productivity, the importance of caring for participants or subjects (i.e., avoiding harm, integrating valid consent processes, and respecting privacy), and economic compensation.

Integrity and Ethical Principles of Research Practices

Integrity is *the* defining characteristic of being ethical in life. Integrity implies consistency and coherence in the application of both technical and ethical principles and the highest standards of professionalism and rigor (Hiney, 2015). Integrity also involves understanding and following the various legal, ethical, professional, and institutional rules and regulations in relation to the activity at hand (Shamoo & Resnik, 2015). Hence, integrity is not only a question of the theoretical acceptance of certain values but also the challenge to put such principles into practice (Resnik, 2011).

Although integrity is considered the cornerstone for ethics in research, it is typically accompanied by six broad values that also shape the professional conduct of the researcher. We discuss them next.

Honesty

Integrity implies acting with honesty. Greater access to well-grounded knowledge serves society and researchers should aim to disseminate their knowledge to society. In this regard, the professional principle of honesty encompasses disclosure, transparency, and confidentiality.

To begin with, researchers should disclose information related to the research process, such as methods, procedures, techniques, and findings. In this manner, society benefits from this provision of information that stimulates academic debate among interested researchers and the public at large (Drenth, 2012; Social Research Association, 2003). Obviously, the publication of this information must be done transparently, without confusion or deception, drawing a clear distinction between their professional statements and any comments made from a personal point of view (Resnik & Shamoo, 2011). This implies with impartiality.

Finally, the dissemination of this information must respect confidentiality, an especially pertinent issue given the ongoing digitalization of society. There are now new forms to disseminate research findings, including online, open-source, and open-access publishing. These new channels provide greater opportunities to share the research, but they also threaten the protection of data relating to the privacy of individuals. Research data cannot be made public without authorization, except for cases where withholding the information might be detrimental to a greater good. For instance, if maintaining a confidentiality agreement were to facilitate the continuation of illegal behavior, such an agreement should be disregarded. Imagine a research biologist employee of a tobacco company who discovers that her firm deliberately works on illegally increasing the addictiveness of the cigarettes but is bound to secrecy due to a confidentiality agreement.

Objectivity

A second principle related to integrity is objectivity (i.e., the use of appropriate methods in research). Researchers must draw conclusions from a critical analysis of the evidence and communicate their findings and interpretations in a complete and objective way (Resnik & Shamoo, 2011). Nevertheless, since the selection of topics can reveal a systematic bias in favor of certain cultural or personal values, researchers should avoid methods of selection designed to produce misleading results or misrepresenting findings by commission or omission (Hammersley & Gomm, 1997).

As an ethical-professional principle, objectivity concerns the need to reach and communicate findings to broaden and enhance knowledge (Drenth, 2012; OECD, 2007; Sutrop & Florea, 2010). Researchers must have pure intentions about the scientific purpose of research, even when they are tempted to act independently of the interests of funders or donors, who may try to impose certain priorities, obligations, or prohibitions (Social Research Association, 2003).

Objectivity affects all aspects of the research process, including experimental design, data analysis, data interpretation, peer review, personnel decisions, grant writing, and expert testimony. Objectivity must also withstand the threat of possible conflicts of interest on financial, political, social, and religious grounds (Israel, 2014). In such cases, researchers must consult ethics protocols (British Psychological Society, 2021). If a conflict of interest of any kind is present, the researcher has a duty to disclose it immediately.

Accountability

According to Mulgan (2000), accountability is “being called to account to some authority.” Thus, the concept implies both a form of social interaction according to which being accountable signifies that a person responds, rectifies, and accepts sanctions; and the rights of authority (i.e., a superior instance that has the right to demand answers and to establish and execute policies of control). This accountability applies to the entire research group. Therefore, researchers are not only accountable for their own research projects but also for the staff and colleagues that are working under their authority. For instance, senior researchers are responsible for and can be held accountable for the use of data that was collected with bias, or without rigorousness, or simply collected negligently by junior staff.

Accountability also covers administrative actions that are transparent and verifiable and that disclose the intents and purposes of the professionals who undertake them. The meaning of accountability is not limited to expressions of honesty or good intentions; accountability means being able to justify that one has acted in an honest and well-intentioned way. Hence, researchers ought to keep a complete record of the research project, in such a way that others may verify and/or replicate the work. Similarly, researchers must take responsibility for their contributions to any publications, funding applications, reports, and other presentations or communications relating to their research (Resnik & Shamoo,2011).

Authenticity

Authenticity in research means to present the findings objectively and rigorously. Hence, authenticity is violated when research findings are misrepresented through incorrect information or confusing data representations, such as false charts and graphics. The aim of misrepresentation is often to gain an advantage from new scientific discoveries: rewards for publication, workplace promotion, a boost in professional prestige, etc. Misrepresentation runs counter to the spirit of research practice (OECD, 2007) and takes on three general forms: fabrication, falsification, and plagiarism (Hiney, 2015; Wible, 2016).

Fabrication

Fabrication is the invention of data or results that are presented as real to prove a working hypothesis (Drenth, 2012). For instance, claims based on incomplete or assumed results is a form of fabrication. This practice is never merely a matter of negligence since it is almost always intentional and fraudulent, and any fabrication of findings is normally considered a serious offence in the science community and society in general.

Falsification

Falsification is the negligent or fraudulent manipulation of existing data to achieve a result that might be expected from the research process (e.g., changing or omitting research results and data to support hypotheses, and claims). The seriousness of falsification lies in the fact that it involves false claims about information that may be relevant for scientific research (Drenth, 2012) rather than whether it satisfies special or specific interests.

Plagiarism

Plagiarism is the appropriation of a person's work, results, processes, etc. (Hiney, 2015), without giving credit to the originator, even if done unintentionally. Plagiarism, therefore, is a form of theft that undermines the integrity of the scientific community and the status of science itself. The problem of plagiarism is particularly acute in relation to citations and acknowledgments in publications. Researchers ought to acknowledge the names and titles of all those who contributed in a significant way to the research project (the European Code of Conduct for Research Integrity), including editors, assistants, sponsors, and others to whom the criteria of authorship do not apply (Resnik & Shamoo, 2011). Guest authorship and ghost authorship are not acceptable because it means giving undue credit to someone. The criteria for establishing the sequence of authors should be agreed by all, ideally at the start of the project. Respect for all aspects of intellectual property is especially important in this regard: patents, copyright, trademarks, trade secrets, data ownership, or any other kind of property in science (Shamoo & Resnik, 2015). Therefore, researchers should also cite all sources and make sure to avoid self-plagiarism, duplication of their own work, or the publication of redundant papers.

Compliance

Compliance is the attitude of attentiveness and respect for rules and regulations, be they the laws of a country, the code of conduct in an organization, trade union conditions, or the norms applicable within scientific associations. Compliance involves ongoing awareness of new rules and regulations governing professional practice (Caruth, 2015; Resnik & Shamoo, 2011; Social Research Association, 2003) in the country or countries where the research is carried out (OECD, 2007). Without such due diligence, unjustifiable lapses may occur through negligence (Sutrop & Florea, 2010).

In situations where others may have engaged in irregular practices, researchers must immediately disassociate themselves from the situation and work toward correcting and redressing the problem. Whistleblowing – the official reporting of such malpractice – is often the only way of putting a stop to the irregular situation. Researchers ought to keep the relevant authorities informed of any suspected inappropriate research conduct. This includes wrong practices such as fabrication, falsification, plagiarism, or any other forms of malpractice that undermines the reliability of research. In particular, this principle should motivate researchers to avoid negligence, incomplete acknowledgment of authors, a lack of information about contradictory data, and/or the use of dishonest methods of analysis (Resnik & Shamoo, 2011).

Care

The last principle is care. This is something which is often highlighted through its absence, when professionals cause harm to third parties, through negligence or intentionally. The former does not imply an intention of causing harm or deceit, but it means that a person does not fulfill what is expected from a reasonable and prudent professional. Additionally, it is generally understood that professionalism not only means making the correct decisions but also to remain informed about the current technical and regulative matters and procedures that society, communities, institutions, and public authorities demand from professionals and that demonstrate the needed care for individuals and society.

The concern for negligent malpractice has mostly been associated with clinical research, especially when drugs and medical procedures can cause damage during experimental trials or treatments. Nevertheless, social and behavioral science researchers are not excluded from the risk of acting negligently (Kaźmierska, 2020). Research processes, such as conducting surveys, can cause a subject to be psychologically affected, maltreated, or aggrieved. Similarly, the disclosure of information and findings can cause reputational damage and create a social stigma associated with communities, institutions, ethnic groups, poor people, etc. or it can mislead society when sensitive information about regular activities, such as health, violence, consumption habits, and political participation, is incorrect or made public with bias.

When social or behavioral scientists fail to safeguard the information provided by their subjects, they are also neglecting their duty of care. The European Council

(2016) identified the importance of avoiding negligence in data management by caring for "processing of personal data to the extent strictly necessary and proportionate for the purposes of ensuring network and information security, i.e. the ability of a network or an information system to resist, at a given level of confidence, accidental events or unlawful or malicious actions that compromise the availability, authenticity, integrity and confidentiality of stored or transmitted personal data."

Moreover, negligence also often occurs during research processes and projects aimed at gathering and analyzing information about violence, abuse, drugs, or psychological conditions. Thus, for researchers to verify whether subjects are in need of professional help and intervention, the research process itself must be conducted responsibly, facilitating information and access to assistance for those subjects in need. The risk of negligence is not only evident when subjects might be harmed but also when assistance is not provided to research subjects or other third parties who already find themselves in harmful situations. Taking care implies that the research activity must strive for the common good as much as it can.

The Context of Professional Research

Scholars need to be aware of the professional ecosystem of research within which they operate. Hence, it is important to reflect on a second criteria that guides how social and behavioral research is ethically conducted by considering the related third parties that constitute the research ecosystem (OECD, 2007). These third parties include universities, research centers, government agencies, the community of scholars, journal editors, publishers, the general public, and research participants.

Research Institutions

Research institutions should develop a strategy to encourage researchers to make positive and quality contributions with societal relevance (Social Research Association, 2009) while creating and providing an ethical culture of scientific work (Drenth, 2012). To do so, institutions must consider two main principles. The first is to narrow the gap between research and practice. Since research is primarily evaluated by its placement in elite journals and its impact on subsequent research, its application to real-world problems is often minor. Universities and research centers usually rely on the impact factor of the journal to determine impact, and journals often favor novelty over cumulative insight (Davis, 2015). The second concern is the quality of the research itself. Academic evaluation systems often encourage quantity over quality (Gupta, 2013), and novelty over replicability, resulting in little cumulative knowledge (RRBM, 2017).

Moreover, institutions must be cognizant of the fact that responsible research is about both useful and credible knowledge. Therefore, institutions ought to appreciate the obligations that researchers have to society at large, research participants, colleagues, and other contributors. Thus, tenure should assess the reliable incremental knowledge as well as the novelty along with its potential for scholarly and societal impact. For the same reason, funding agencies and government agencies should broaden the criteria for funding decisions to include societal impact in addition to intellectual merit (RRBM, 2017).

This commitment to integrity should involve clear policies and procedures, training, and mentoring of researchers. Institutions should also demonstrate integrity in their activity and avoid any form of malpractice. Such oversight may be carried out by ethics committees, departments or teams, or by means of codes of conduct (Social Research Association, 2003). In addition, research centers, journals, and any other research-related professional associations or organizations should have clear procedures on how to deal with accusations of ethical failures or other forms of irresponsible malpractice (Resnik & Shamoo, 2011).

The responsibility of researchers to their employer or institution includes accountability for financial matters, how the project marks progress in knowledge, whether or not it meets the legitimate interests of the institutions, and if it satisfies the relevant economic, cultural, and legal requirements (Iphofen, 2011; Sutrop & Florea, 2010). This responsibility may be complemented with an explanation of how the research project contributes to the development of skills, competencies, and knowledge across the institution (Social Research Association, 2003).

The Scientific Community

Research is a collective endeavor that normally involves cooperation among a number of researchers as well as the sharing of data and findings with the rest of the scientific community. However, the idea of scientific community is a broad one and extends beyond the immediate research team (Sutrop & Florea, 2010), encompassing all national and international researchers (Drenth, 2012). Thus, the professional activity of the researcher must consider the overall purpose of the scientific community, including its reputation and prestige (Hansson, 2011). This commitment involves responsible teamwork, authorship, peer reviewing and editing, and mentoring, alongside the work of research ethics committees, as detailed below.

Responsible Teamwork

The relationships between members of a research team should contribute to the professional development of each researcher – neither coming at a cost to other team members nor limiting their professional growth (Social Research Association, 2009).

Responsible Authorship

Responsible authorship requires the publication of quality scientific research, the enhancement of scientific knowledge, meeting the needs of the funding institution, and ensuring that the findings published are relevant to society. A key aspect of this is acknowledgment: the right of co-authors to be recognized as such and receive whatever benefits may be due to them as a result (Drenth, 2012). Researchers should ensure that decisions about authorship are made in a fair and transparent way, acknowledging every relevant contributor, according to the established conventions of the discipline.

Responsible Peer Reviewing and Editing

The responsibilities of anonymous reviewers and journal editors include ensuring high scientific standards in publications while advancing knowledge. In particular, reviewers and editors should help researchers improve their research by providing them with recommendations and further readings. Reviewing and editing should be also carried out in accordance with objective criteria and be attentive to possible conflicts of interest. For instance, a reviewer should decline to review a work if s/he knows the authors or if that work could compete with his/her own work (Social Research Association, 2009). It would be unethical for reviewers to make use of any materials submitted to them, for their own purposes or the purposes of any third parties, without the express permission of the authors. See Chapters 33 and 34 in this volume for a further discussion.

Responsible Mentoring

Responsible mentoring implies training new researchers, PhD candidates, postgraduate students, or postdoctoral scholars to make them capable of contributing in a significant way to the scientific community (Drenth, 2012) and helping them progress in their academic careers (OECD, 2007). For instance, the mentors should train mentees to present their findings at conferences and in ethical ways of conducting research.

Research Ethics Committees

The goal of data management is to effectively administer any information that may be of use to current or future scientists (Drenth, 2012). Many research data – even sensitive data – can be shared ethically and legally if researchers employ strategies of informed consent, anonymization, and controlling access to data (UK Data Service, 2021); good data management is nothing less than an ethical duty for the scientific community.

Accordingly, research ethics committees should design policies in relation to the availability of data and the criteria governing appropriate use. This includes advising researchers how to store data in a secure and protected manner, share, and make it available for reuse and complying with the requirements of data protection legislation and best practices concerning confidentiality. In this way, the role of research ethics committees is to protect the safety, rights, and well-being of research participants and to promote ethically sound research. In addition, confidentiality agreements should always be respected, irrespective of whether they were originally established for the purposes of previous research projects, except for cases where withholding the information might be detrimental to a greater good.

Managing Research Productivity

Research productivity is often measured by publishing in academic journals and having an impact on the scientific community. Hence, when universities and research centers have professional managers responsible for the allocation of resources and career development, the systematization of productivity standards gains importance for organizational policy making, strategy, faculty recruiting, student admission, human resource management, and other aspects of work life. The challenge for many of these institutions is to ensure accurate and reliable quality standards.

The practice of qualifying publications began in 1927 when two chemists at Pomona College wrote an article in *Science* proposing that librarians could use data about citation rates to select appropriate journals for a small library collection. This idea was successful, and universities and research centers began to use these rates to allocate resources to both librarians and several other academic activities (e.g., funding and academic appointments, developing faculty policies for tenure positions, and career development). Research productivity has also been assessed in terms of impact within the scientific community. This practice started with the introduction of the impact factor developed by Eugene Garfield in 1975, who founded the Institute for Scientific Information (ISI) that provides the Journal Citation Report (JCR). The impact factor is used as a proxy for the relative importance of a journal (journal impact factor or JIF) by reflecting the average annual number of citations of an article. Hirsch (2005) proposed the metric for having a parameter for personal research productivity, using the h-index (i.e., the number of cites registered for a period).

However, the introduction of these standards for research activity has created a prolonged controversy in the academic community. For instance, the DORA declaration (https://sfdora.org/) in which more than 19,000 researchers and institutions worldwide recommend not using "journal-based metrics as a surrogate measure of research quality." Indeed, metrics for assessing productivity make it difficult to value the real contribution, originality, insights, and influence of a scholar's research (Norris, 2021). Moreover, if the allocation of economic resources, funding, grants, or any other form of economic benefits depends solely on the impact factor of the research, an ethical conflict can easily arise. In other words, research productivity can stray from the main goal of research – advancing social knowledge and public good – and inadvertently promote results-seeking practices at the expense of the real importance of academic activity.

Dealing with Human Participants

The treatment of human participants may be the most significant ethical concern in the social and behavioral sciences. Current codes of research conduct largely came about because of the findings of the Nuremberg trials concerning the medical experiments performed by the Nazis (Kopp, 2006; Sutrop & Florea, 2010) and other medical, surgical, clinical, and psychological experiments conducted in the United States during the twentieth century. Similarly, the National Commission for the Protection of Human Services of Biomedical and Behavioral Research in the United States published the Belmont Report in 1974 in response to the unethical treatment of patients who participated in a medical study. This growing recognition of the dignity of the human person has culminated in a broad consensus that research must always respect the dignity of each and every human being, living or dead (Wilkinson, 2002).

Dignity has many meanings in common usage (Dan-Cohen, 2012; Sulmasy, 2008), but principally refers to the intrinsic worth or value of every human being that distinguishes him/her from any other being, and as such, merits respect (Sison et al., 2016). Such worth or value is often associated with the capacity for reason and autonomy or "self-determination" through free choice. It also implies the need for consensus or mutual recognition among fellow human beings. In short, dignity refers to a preeminent social position that has come to be attributed universally to all human beings (Dan-Cohen, 2012).

In general, the protection of the personal dignity of human subjects behooves research projects to take the following issues into account: avoiding harm, obtaining valid consent, respecting privacy, and economic compensations for participants, which are considered below.

Avoiding Harm

Social and behavioral research does not expose human subjects to as much harmful effects as other forms of scientific research (Bell & Bryman,2007). Nevertheless, it is vital to ensure that no research project involves serious physical, psychological, or moral harm or injury to any participant (Barret, 2006; General Assembly of the World Medical Association, 2014; Sutrop & Florea, 2010; Social Research Association, 2003).

Obtaining Valid Consent

The dignity of the person precludes any form of coercion obliging an individual or individuals to participate in a research project. Participation must be freely undertaken based on an informed decision (Israel, 2014), by giving explicit consent in accordance with a clear protocol, the law, and the culture of the participants (Sutrop & Florea, 2010; see also Chapter 10 in this volume). The latter condition includes the responsibility to offer a complete description of the project, including all relevant research details required to give truly informed consent. As explained by Villaronga et al. (2018), consent should be given by a clear affirmative act establishing a freely given, specific, informed, and unambiguous indication of the data subject's agreement to the processing of personal data, such as by a written statement, including by electronic means or an oral statement.

This informed consent implies that the participant has to understand the relevant information about the process and goals. The participation also has to be voluntary. In other words, the consent is given freely and not as a result of coercive pressure (real or perceived). It must also be competent, meaning the consent must be given by somebody capable, by virtue of their age, maturity, and mental stability, of making a free, deliberate choice (Houston, 2016; Macfarlane, 2010).

The need for informed consent does not preclude the possibility of addressing those who lack the competence to offer free and informed consent. In fact, excluding children younger than the required age to give consent, those who have not reached the age of reason, people with learning or communication difficulties, patients in care, people in custody or on probation, or people engaged in illegal activities (such as drug abuse), may in itself constitute a form of discrimination (Social Research Association, 2003). This kind of research often yields results that may contribute to bettering the quality of life of these groups of people since the effectiveness of public policies enacted in relation to them depends on previous research of this kind. In these cases, informed consent should be given either by parents or by legal guardians. Such third parties must give valid consent, under the same conditions as a standard expression of consent, and confirm that there are no conflicts of interest (Social Research Association, 2003). In addition, depending on the potential risks to the participants, it may be necessary to obtain advice and approval from an independent ethics committee (General Assembly of the World Medical Association, 2014).

Respecting Privacy

The principle of personal dignity also requires that the legitimate privacy and decorum of research subjects is safeguarded, regardless of the value of the research being undertaken and/or the potential of the new technologies now deployed for research purposes (Social Research Association, 2003). This condition sets limits on the information that may be sought about a person for the purposes of any research project, especially if s/he has not consented to disclosing every detail about their personal life (Barret, 2006).

We must look no further back than 2017 to find a cogent example of the violation of personal consent and privacy. The Facebook and Cambridge Analytica exposé triggered a public debate on the responsibility of social media and internet companies and their third-party partners regarding the use of personal data and the degree of consent that users willingly give. Cambridge Analytica is a political consulting firm, which combines data mining, data brokerage, and data analysis alongside strategic communication for electoral campaigning purposes. It managed to collect data and build profiles of millions of Facebook users using sources such as demographics, consumer behavior, internet activity, and, most worryingly, by collaborating with others who were given user data by Facebook under the pretext of academic research. In March 2018, *The New York Times* and *The Observer* reported that Cambridge Analytica, however, used this personal information for its commercial service offering to influence the outcomes of the 2016 US elections and the Brexit referendum, without the users' permission or knowledge and without the permission from Facebook. In sum, the principle of privacy limits the use researchers may make of the data they collect because such information is provided to a given researcher or research institution, but not granted to the scientific community or society at large, even though it might be beneficial to them. Therefore, access to research data must be subject to tight control (Bell & Bryman, 2007). Anonymization, for instance, is a valuable tool that enables the sharing of data while preserving privacy. The process of anonymizing data requires that identifiers be changed in some way by being removed, substituted, distorted, generalized, or aggregated.

Economic Compensations for Participants

According to Grady (2005), paying participants in a research project has several practical benefits, especially if it enhances recruitment and overcomes the risk of inertia, lack of interest, and financial barriers to those who, without any stipends, would not be able to participate. By awarding economic benefits to participants, the research goals can also be broadened along with the racial, gender, ethnic, and social diversity.

However, using financial incentives raises important ethical concerns. Indeed, when payment policies are not correctly understood and applied, they can provoke bias and a lack of objectivity and affect the research findings. Several authors have explained why subjects must not be paid or in which circumstances some forms of compensation cannot be accepted. For instance, Dickert et al. (2002) maintained that payment must be considered in a scheme of justification, formula, and restriction. Justification means "what are researchers paying for" and this can be described as incentives, time, travel and food, inconvenience, risk, etc. There are also different formulae and methods for remunerating research participants, such as cash, pro rata payments, or by providing other benefits or products (e.g., discount cards or academic benefits for students). Participants should also meet certain conditions to be paid, such as having completed the research, or belong to a specified racial, ethnic, or social group.

Moreover, it is important to establish and disclose the process to determine and allocate the amounts for economic compensation (Grady, 2005). In this regard, a research project can simply use market standards, or in other words, pay what research participants are willing to accept. Alternatively, the payment amount can be based on the standardized hourly wage that might result in different rates, depending on the occupation and normal salary of each subject. These market-related agreements and salary schemes could also incorporate additional incentives such as completion bonuses and escalation fees. Another form of determining the payment amount is to compensate for any inconveniences and reimburse research participants for expenses (e.g., travel and meals). Since these are variable costs, it would be good practice to cap expense claims to an appropriate limit and to consider an insurance policy to protect participants against any foreseeable risks.

Ethical concerns also arise when participants are paid for their involvement in research, including avoiding bias and avoiding coercion. Bias comes from the fact that (i) participants might not be providing the same information if they were not paid and (ii) the aim of the research project would have included willing participants even though they are not being paid. Regarding coercion, it is important that participants in the research project are not indirectly pressured to partake in the project. Coercing can be seen in those cases in which participants are in dire financial need or in the case of students whose professors or advisers are conducting the project. For McNeill (1997), inducing subjects to participate by being paid may invalidate informed consent, especially when these subjects are poor.

Digital Behavioral Social Science Research

Computational social science is an instrument-based discipline that enables the observation and empirical study of the social phenomena associated with the human–computer interaction in society. This interaction includes individual cognition, decision-making, group dynamics, organization and management, and societal behavior in local communities (Kirilova & Karcher, 2017). As explained by Edlund et al (2017), research projects developed online and in digital platforms can cause bias, especially when participants interact with each other. Thus, some ethical concerns derive from the fact that such digital resources can affect responsible research practice and scientific knowledge. Nevertheless, the moral concern for research in a digital context is mainly focused on the fact that participants are using and providing, whether explicitly or not, sensitive personal information.

While it is true that significant progress has been made regarding data protection, it is also the case that there is an increasing volume of information pertaining to companies, non-governmental organizations, and government bodies available via the Internet that may be used for the purposes of research but without the relevant authorizations from the proprietors of those networks and/or websites (Hansson, 2011; Social Research Association, 2003, 2009). This includes the study conducted by a Facebook scientist and two academics, aimed at testing whether emotional contagion occurs between individuals on Facebook. In this project, researchers manipulated Facebook's newsfeed by showing fewer positive posts to examine if this would correlate with greater user expressions of sadness. Among other concerns, this experiment raised issues related to the difficulty of knowing if Facebook's users were aware that they gave consent to the use of their interactions on the social network for research purposes (McNeal, 2014).

The Internet is becoming a global resource of data (Kaźmierska, 2020). Hence, although the ethical concerns that arise in research carried out via the Internet are similar to those that emerge in other forms of research, the specific characteristics of the virtual world pose a unique challenge regarding informed consent, confidentiality, and the security of data transmission (Grinyer, 2009). Accordingly, qualitative researchers must consider data-archiving their reanalysis and determining the boundary for creating qualitative "big data" (i.e., "big qualidata"). Thus, computational social science research activities must adhere to specific principles and values that stem from the basic right of users to keep their personal data private. As explained by Kirilova and Karcher (2017), digital formats of text, audio, and video contribute to qualitative data becoming more readily available and increasingly easy to distribute. This has led to a rapidly growing interest in managing and sharing qualitative data that has increased the risk of private human subject data inadvertently being made public. In this regard, the regulation provided by the European

Parliament in 2016 (European Council 2016) has identified the following considerations in which participants might not necessarily have given consent for the use of their personal information in research activities.

Subjects' Right to Rectification

When subjects are holders of their personal information, even when it is shared, they are entitled to rectify inaccurate or incomplete data by providing supplementary statements or through other remedial actions.

Subjects' Right to be Forgotten

The right to be forgotten or *data deletion* aims to reduce the public accessibility and use of private information if the users have not declared this information as public or if they have limited the information to a specific use (e.g., research purposes). This right also has a retroactive effect, meaning that the historical data of a user is under the right to be restricted and/or erased. This right applies to explicit sharing of information as well as any other information made public by individuals, except in cases of public interest.

According to the European Council (2016), "The data subject shall have the right to obtain from the controller the erasure of personal data concerning him or her without undue delay and the controller shall have the obligation to erase personal data without undue delay." Moreover, erasing data means more than just deactivating or impeding the data being shared publicly; it also implies eliminating data from the digital storage. According to Villaronga et al. (2018), the data deletion right demands that data controllers act without undue delay when shared data are no longer necessary for the original purpose under which data were given.

Subjects' Right to Restriction of Processing

The use of data must be limited to the individual or corporate/institutional entity with whom the subjects have explicitly agreed to share information with and limited to the agreed and specific purpose. This means that other individuals or institutions cannot use data for any other purpose other than what was originally agreed with participants. Accordingly, storing data by institutions or corporations does not permit them to use these data for any different purpose than the original, and it shall not be used by any other data controller who does not belong to that institution or corporation. Consequently, data controllers are bound by what is defined as *data minimization policies* (i.e., to work with only information or data deemed strictly necessary or important in a research project; Villaronga et al., 2018).

Participants' Right to Data Portability

This right safeguards the fact that, even when personal information has been provided to a controller, the participant does not lose his/her right to transmit that information to another controller.

Participants' Right to Object and Automated Individual Decision-Making

Finally, participants have the right to object to their personal information being processed for scientific or historical research purposes or statistical purposes, unless the processing is necessary for the performance of a task carried out for reasons of public interest. This right makes sure that researchers uphold accountability and transparency during the research process, respecting anonymity and ensuring that participants are aware of the research activity and their right to withdraw personal information.

Self-Assessing a Research Project

When we accept that complying with ethical standards is not only a matter of adhering to bureaucratic procedures that many researchers find unnecessary, overwhelming, and distracts them from the essence of their academic work, it becomes useful and important to posit some general questions that might guide us to grasping and assessing the ethics of our research projects. Based on the principles of honesty, accountability, compliance, care, and authenticity, we recommend using the following questions as 'friendly reminders' during the different stages of the lifecycle – before, during, and after a research project.

Honesty

- Does this research project advance science and knowledge as a public good?
- Does this research project aim at policy making in private institutions or government?
  - If so, how is the policy described for this project?
- What is the dissemination strategy and publication goal for this project?
- What are the specific and general aims of this project?
- Are the methods, procedures, and techniques of this research project explicit for participants, committees for research ethics, funding boards, etc.?
- Who are the academic stakeholders of this project? Are all of them informed?
- Who is entitled to receive the information about the project and its findings?
- Can the findings be made public? If so, when, and how?
- Can third parties be affected or stigmatized by sharing information about this project and its findings?
- How is the dissemination strategy defined for the scientific community and society?
- Who is responsible for providing information associated with the project before it is made public?
- Are all participants informed of their non-disclosure agreements?
- Is every member of the research team informed about their authorship or co-authorship in the planned publications and research outcomes?
- Are trademarks, licenses, and copyright considered and agreed with related institutions, organizations, and constituencies?

Accountability

- Have procedures been defined to ensure that data backups are made that can be made available to third parties and authorities when requested?
- Who is accountable to whom for the project?
- Are private funding and public grants explicitly stated for this project? How and where is the information about funding made public?
- How are the required documents to be delivered to and accepted by third parties when needed?
  - Signed contract and agreements for each member of the research team.
  - Signed informed consent for participants.
  - Disclosure of conflict of interest for research team members and participants
- Has this project identified the risks and potential harm (physical, sociological, spiritual, political, reputational, etc.) for participants and research team members?
- Has the privacy of participants and research team members been respected?
- Are participants paid?
  - If so, what kind of payment agreement has been established for this project (money, benefits, pro rata, etc.)?
  - How is the amount paid determined (market practice, wage, reimbursement, etc.)?
- Has this project avoided coercion for paid participants? How?

Authenticity

- If applicable, has this project avoided fabrication or data bias?
- Has the responsible researcher avoided risks of plagiarism associated with other researchers, staff, pollsters, interviewers, etc.?

Compliance

- If applicable, does this project comply with the norms and regulations stipulated by the committee for research ethics?
- Have the leading researchers identified, if applicable, legal risks associated with this project, especially in terms of defamation and negligent submission of public information?

Care

- Does this research project aim at exploring situations of violence, abuse, health issues, antisocial tendencies, etc.?
- Have the survey staff been prepared to offer alternatives of assistance to participants or have they followed the protocols to inform authorities without violating the privacy of the participants?
- In the case of surveys, do the questions avoid sensitive words, offensive categorizations, social stigmas, etc.?
- Has the strategy for the public dissemination of findings (especially charts, press briefings, executive briefs, social networks, etc.) been assessed by a committee or a research group to avoid providing misleading information and facilitating wrong/partial interpretation of data by society and public authorities?
- Have the participants of the research project provided valid consent in their full capacity?

  - If not, how has this project obtained consent from the legal guardians of participants?
  - Has valid consent delivered by participants or legal guardians been signed with legal and valid signatures, digital or written?
- How are the rights to rectification, erasure, restriction of processing, portability, and objection being protected? (This is especially important for participants in digital social science research projects.)
  - If applicable, are these rights protected even when information used for research is public and has not been given by informed consent?

Conclusion

Contributing to a better world is the ultimate goal of science (RRBM, 2017). Social and behavioral science research can live up to this duty if it continues to hold the values outlined in this chapter in the highest esteem. Ethics helps researchers to carry out their research honorably, honestly, objectively, and responsibly toward the overarching goal of enhancing our understanding of human behavior in social contexts. On the other hand, given that research is a collective endeavor involving a significant degree of collaboration and cooperation, ethics facilitates the progress of science by underscoring key values such as teamwork and trust, responsibility, generosity, respect, fairness, and authorship among others that we discussed in this chapter.

In addition, ethics ensures that research meets the aims and needs of funding institutions, respecting both the legitimate interests of those bodies and the broader interests of society. In this way, ethically grounded research generates knowledge that may be useful and valuable to society, publishing it in a transparent way, wholly respectful of the safety, privacy, confidentiality, and dignity of the participants. Ethical research contributes to civil society as well as to the funding bodies that finance research studies.

Finally, the service and benefit afforded by rigorous research means that methods, procedures, techniques, and findings contribute to refining and enhancing knowledge and these are made available to the scientific community, thus furthering the academic endeavor. These are some of the reasons that account for the growing importance of the field of research ethics. We believe that it is imperative that ethics is included in the curricula for the pedagogical development of anyone pursuing work in professional research. Education in ethical research will lead to a greater understanding of ethical standards, ethics policies, codes of conduct, and, above all, how sound ethical judgment, decision-making, and practical wisdom among researchers can continue to be fostered.

References

Barrett, M. (2006). Practical and ethical issues in planning research. In G. Breakwell, S. M. Hammond, C. Fife-Schaw, & J. A. Smith (eds.), *Research Methods in Psychology,* 3rd ed. (pp. 24–48). Sage.

Bell, E. & Bryman, A. (2007). The ethics of management research: An exploratory content analysis. *British Journal of Management*, *18*(1), 63–77.

Brandt, A. M. (1978). Racism and research: The case of the Tuskegee Syphilis Study. *Hastings Center Report*, *8*(6), 21–29.

British Psychological Society (2021). *Code of Human Research Ethics*. British Psychological Society.

Caruth, G. D. (2015). Toward a conceptual model of ethics in research. *Journal of Management Research*, 15, 23–33.

Dan-Cohen, M. (2012). Introduction: Dignity and its (dis)content. In J. Waldron (ed.), *Dignity, Rank, and Rights* (pp. 3–10). Oxford University Press.

Davis, G. F. (2015). Editorial essay: What is organizational research for? *Administrative Science Quarterly*, *60*(2), 179–188.

Dickert, N., Emanuel, E., & Grady, C. (2002). Paying research subjects: An analysis of current policies. *Annals of Internal Medicine*, *136*(5), 368–373.

Dingwall, R., Iphofen, R., Lewis, J, Oates J., & Emmerich, N. (2017). Towards common principles for social science research ethics. A discussion document for the Academy of Social Sciences. In R. Iphofen (ed.), *Finding Common Ground: Consensus in Research Ethics Across the Social Sciences* (ch. 10). Emerald Publishing Limited.

Dotterweich, D. P. & Garrison, S. (1997). Research ethics of business academic researchers at AACSB institutions. *Teaching Business Ethics*, *1*(4), 431–447.

Drenth, P. J. (2012). A European code of conduct for research integrity. In T. Meyer & N. Steneck (eds.), *Promoting Research Integrity in a Global Environment* (pp. 161–168). World Scientific Publishing.

Edlund, J. E., Lange, K. M., Sevene, A. M., et al. (2017). Participant crosstalk: Issues when using the mechanical Turk. *Tutorials in Quantitative Methods for Psychology*, *13*(3), 174–182.

ERC (European Research Council) (2021). ERC work programme 2021. Available at: https://erc.europa.eu/content/erc-2021-work-programme.

European Council (2016). Regulation (EU) 2016/679 of 27 April 2016.

General Assembly of the World Medical Association (2014). World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. *The Journal of the American College of Dentists*, *81*(3), 14.

Grady, C. (2005). Payment of clinical research subjects. *The Journal of Clinical Investigation*, *115*(7), 1681–1687.

Grimes, J. M., Fleischman, K. R., & Jaeger, P. T. (2009). Virtual guinea pigs: Ethical implications of human subjects research in virtual worlds. *International Journal of Internet Research Ethics*, *2*(1), 38–56.

Grinyer, A. (2009). The anonymity of research participants: Assumptions, ethics, and practicalities. *Pan-Pacific Management Review*, *12*, 49–58.

Gupta, A. (2013). Fraud and misconduct in clinical research: A concern. *Perspectives in Clinical Research*, *4*(2), 144.

Hammersley, M. & Gomm, R. (1997). Bias in social research. *Sociological Research Online*, *2*(1), 7–19.

Hansson, S. O. (2011). Do we need a special ethics for research? *Science and Engineering Ethics*, *17*(1), 21–29.

Hiney, M. (2015). *Research Integrity: What It Means, Why Is So Important and How We Might Protect It*. Science Europe.

Hirsch, J. E. (2005). An index to quantify an individual’s scientific research output. *Proceedings of the National Academy of Sciences*, *102*(46), 16569–16572.

Houston, M. (2016). The Ethics of Research in the Social Sciences: An Overview. University of Glasgow. Available at: https://dafre.rutgers.edu/documents/Articles_Ethics_research_%20social_sciences.pdf.

Iphofen, R. (2011). Ethical decision making in qualitative research. *Qualitative Research*, *11*(4), 443–446.

Israel, M. (2014). *Research Ethics and Integrity for Social Scientists: Beyond Regulatory Compliance*. SAGE Publications.

Kanungo, R. N. (1992). Alienation and empowerment: Some ethical imperatives in business. *Journal of Business Ethics*, *11*(5–6), 413–422.

Kaźmierska, K. (2020). Ethical aspects of social research: Old concerns in the face of new challenges and paradoxes. A reflection from the field of biographical method. *Qualitative Sociology Review*, *16*(3),118–135.

Kirilova, D. & Karcher, S. (2017). Rethinking data sharing and human participant protection in social science research: Applications from the qualitative realm. *Data Science Journal*, 16, 43.

Kopp, O. (2006). Historical review of unethical experimentation in humans. *Ethics in the Professions*, 2. Available at: https://w.astro.berkeley.edu/~kalas/documents/ethics/2007facultyproceedings_g5small.pdf#page=16.

Lock, S. (1995). Research ethics. A brief historical review to 1965. *Journal of Internal Medicine*, *238*(6), 513–520.

Macfarlane, B. (2010). *Researching with Integrity: The Ethics of Academic Enquiry*. Routledge.

McNeal, G. (2014). Facebook manipulated user news feeds to create emotional responses. *Forbes,* June 28. Available at: www.forbes.com/sites/gregorymcneal/2014/06/28/facebook-manipulated-user-news-feeds-to-create-emotional-contagion/?sh=2e007a1d39dc.

McNeill, P. (1997). Paying people to participate in research: Why not? *Bioethics*, *11*(5), 390–396.

Mulgan, R. (2000). ‘Accountability’: An ever-expanding concept? *Public Administration*, *78*(3), 555–573.

Norris, P. (2021). What maximizes productivity and impact in political science research? *European Political Science*, *20*(1), 34–57.

OECD (Organisation for Economic Co-Operation and Development) (2007) Global Science Forum: Best practices for ensuring scientific integrity and preventing misconduct. Available at: www.oecd.org/science/inno/40188303.pdf.

Resnik, D. B. (2011). What is ethics in research & why is it important? *The National*. May.

Resnik, D. B. & Shamoo, A. E. (2011). The Singapore Statement on research integrity. *Accountability in Research*, *18*(2), 71–75.

RRBM (Responsible Research in Business & Management) (2017). A vision of responsible research in business and management: Striving for useful and credible knowledge. Available at: www.rrbm.network/position-paper (accessed February 27, 2018).

Ruyter, K. W. (ed.) (2003). *Forskningsetikk: beskyttelse av enkeltpersoner og samfunn*. Gyldendal akademisk.

Shamoo, A. E. & Resnik, D. B. (2015). *Responsible Conduct of Research*. Oxford University Press.

Sison, A., Ferrero, I., & Guitián, G. (2016). Human dignity and the dignity of work: Insights from catholic social teaching. *Business Ethics Quarterly*, *26*(4), 503–528.

Social Research Association (2003) Ethical guidelines. Available at: https://the-sra.org.uk/common/Uploaded%20files/ethical%20guidelines%202003.pdf.

Social Research Association (2009) Social Policy Association Guidelines on Research Ethics. Availabale at: https://social-policy.org.uk/wp-content/uploads/2014/05/SPA_code_ethics_jan09.pdf.

Sulmasy, D. (2008). Dignity and bioethics. History, theory, and selected applications. In *Human Dignity and Bioethics* (pp. 469–501). The President's Council on Human Dignity and Bioethics.

Sutrop, M. & Florea, C. (2010). *Guidance Note for Researchers and Evaluators of Social Sciences and Humanities Research*. European Commission.

UK Data Service (2021). Research data management. Available at: https://ukdataservice.ac.uk/learning-hub/research-data-management (accessed November 3, 2021).

Villaronga, E. F., Kieseberg, P., & Li, T. (2018). Humans forget, machines remember: Artificial intelligence and the right to be forgotten. *Computer Law & Security Review*, *34*(2), 304–313.

Wible, J. R. (2016). Scientific misconduct and the responsible conduct of research in science and economics. *Review of Social Economy*, *74*(1), 7–32.

Wilkinson, T. M. (2002). Last rights: The ethics of research on the dead. *Journal of Applied Philosophy*, *19*(1), 31–41.

Zimbardo, P. G., Haney, C., Banks, W. C., & Jaffe, D. (1971). The Stanford prison experiment. Available at: https://web.stanford.edu/dept/spec_coll/uarch/exhibits/Narration.pdf.

3 Getting Good Ideas and Making the Most of Them

Christian S. Crandall and Mark Schaller

Abstract

Good research ideas and hypotheses do not just magically exist, begging to be tested; they must be discovered and nurtured. Systematic methods can help. Drawing on relevant scholarly literatures (e.g., research on creativity) and on the published personal reflections of successful scientists, this chapter provides an overview of strategies that can help researchers to (1) gather research ideas in the first place, (2) figure out whether an idea is worth working on, and (3) transform a promising idea into a rigorous scientific hypothesis. In doing so, it provides pragmatic advice about how to get good ideas and make the most of them.

Keywords: Ideas, **Hypotheses, Creativity, Research Methods**

Introduction

Scientific progress occurs through a kind of evolutionary process. Scientists identify innovative new ideas and hypotheses about what might be true, and they use empirical methods to test them (i.e., to eliminate those that fail to meet accepted standards of evidence and to selectively retain those that do; Campbell, 1974; Hull, 1988; Popper, 1963). Both parts of this process are equally essential to scientific progress, but they receive unequal attention within scientific education. Scientists receive enormous amounts of formal training in methods to use and best practices to employ when testing ideas and hypotheses against empirical data. That's good. In contrast, scientists typically receive very little formal training in methods and practices that might help them to identify new ideas and develop new hypotheses in the first place. That's too bad.

Scientific ideas and hypotheses don't just magically exist, begging to be tested. They must be discovered and developed by scientists themselves and communicated coherently to other people in the scientific community. Just as the empirical testing part of the scientific process benefits from strategy and methodological skill, so too does this innovation part of the process. Systematic strategies can be used to increase the likelihood of being inspired with innovative ideas and to determine whether those ideas are worth pursuing or not. It takes both strategy and skill to transform an informal idea into a precise, logically coherent scientific hypothesis.

That is why this handbook includes this chapter. We've designed it to provide systematic methodological guidance – and pragmatic advice – about how to get good ideas and make the most of them.

Strategies for Gathering Ideas and Lots of Them

There is a lovely line in the novel *Of Love and Other Demons* (García Márquez, 1995, p. 56): "Ideas do not belong to anyone . . . They fly around up there like the angels." What you want is for some of those ideas to fly from the sky and grace your brain with inspiration. It's not merely luck; scientists can do things to make it happen, again and again and again.

A first rule of thumb: *At the early stages, don't worry about whether those ideas are good ones or not*. This might seem counter-intuitive because scientific training emphasizes methods to diagnose the rightness or wrongness of ideas. However, that diagnostic work comes later, and it cannot happen until *after* inspiration has occurred. A self-critical mindset is useful when designing studies, when analyzing data, and when drawing conclusions from those data, but it's counterproductive to creativity (Lam & Chiu, 2002.)

To invite inspiration that might be right, savvy scientists allow themselves to be wrong. At this earliest stage of scientific discovery, it is helpful to cultivate a mindset that is open to anything, including good ideas, mediocre ones, and even mistakes. (This rule of thumb is collected with ten more in Box 3.1.)

Box 3.1 Eleven useful rules of thumb for getting good ideas and making the most of them

Getting good ideas

(1) At the early stages, do not worry whether your ideas are good ones or not.
(2) Really good ideas do not often start out as really good ideas.
(3) Expose yourself to diversity; new experiences promote creativity.
(4) Do things that you actually want to do; intrinsic motivation helps.
(5) Inspiration is idiosyncratic; try many things.

Making the most of them

(6) Interact with other people – talk, share, disagree, discuss, and agree.
(7) Ideas with real-life relevance tend to find more people who are interested in them.
(8) If an idea is too obviously true, people might not find it interesting.
(9) Define carefully and precisely an idea's conceptual components and state their relations to each other.
(10) Ideas do not belong to anyone; avoid identifying with "your" hypotheses.
(11) Specify your assumptions explicitly.

This open-minded perspective is encouraged by many philosophers of science. Paul Feyerabend (1975, p. 17) wrote: “Science is an essentially anarchic enterprise: theoretical anarchism is more humanitarian and more likely to encourage progress than its law-and-order alternatives . . . The only principle that does not inhibit progress is: *anything goes*.” One reason why it is okay to adopt an “anything goes” approach to inspiration is because of the communal and self-correcting nature of the scientific process. David Hull (1988, p. 7) reminds us that “Science is a conversation with nature, but it is also a conversation with other scientists.” We are allowed, even encouraged, to introduce ideas of any kind into that conversation because the conversation – the scientific process which follows any act of inspiration – judges those ideas rigorously and can refute those that fail to meet strict standards of evidence. The refutation of flawed conjectures is *essential* to scientific progress (Popper, 1963). As long as you commit in good faith to that rigorous process, there’s little harm to making mistaken conjectures; they are common, inevitable, and can even be useful in unexpected ways. Again, we defer to a philosopher of science, Ilkka Niiniluoto (2019): “scientific theories are hypothetical and always corrigible in principle. But even when theories are false, they can be cognitively valuable.” Any idea – whether “right” or “wrong” – has the potential to help point scientists in the right direction. And if an idea is downright dumb? No problem. Scientists are pragmatic; if an idea is unproductive, it won’t be pursued for long. Successful science is littered with wrong ideas. From Archimedes to Ahmed Zewail, from Ainsworth to Zajonc, every serious scientist has had them. They also had ideas that turned out to be right, and one reason they did was because they were willing to be wrong.

A second rule of thumb: *Really good ideas usually do not start out as really good ideas*. They often start out as vague thoughts, niggling questions, half-baked observations. One of us once started with nothing more than a catchy title. It eventually turned into an extensive, rigorous, multi-study research project (Bahns et al., 2017). The supposedly “catchy” title was never used, as it turned out to be less good than the idea it turned into. Another personal example: A random bit of laughably amateurish musing about infectious diseases blossomed – after conversations and collaborations with many people – into a multi-pronged program of research on the “behavioral immune system” (e.g., Murray & Schaller, 2016), within which dozens of new hypotheses have been generated and tested, with wide-ranging implications for human cognition, human behavior, and human culture.

Simply start with inspiration – even laughably amateurish ones. A promising idea will surely be improved, truly unpromising ones will be discarded, and the scientific conversation will help you sort out which is which.

Cultivating a Receptive Mind

Some people are more creative than others (Feist, 1998), but *everyone* has the capacity for inspiration, and *anyone* can discover useful hunches and hypotheses. To do so, one must be receptive. Research on creativity suggests some strategies that can help you cultivate a receptive mind.

A third rule of thumb: *Expose yourself to diversity*. Young scholars are sometimes advised to narrow their interests or to focus their reading on the restricted range of academic literature that is most directly pertinent to their particular academic discipline. That advice may be well-intentioned, and perhaps even pragmatic in a short-sighted way, but it can inhibit inspiration and cramp creativity. The most creative people in the sciences tend to have interests and skills that transcend disciplinary boundaries (Root-Bernstein & Root-Bernstein, 2004; see also Chapter 32 in this volume). Successful scientists often find inspiration in their non-scientific interests, and their non-academic activities often nourish and serve their academic aspirations. Creativity is fueled by exposure to diverse people, places, activities, and perspectives.

Exposure to diverse cultures fortifies the cognitive foundations of creative thought and enhances innovation (Leung et al., 2008). You may not have to sojourn to a far-away land to benefit (but it can help; Maddux & Galinsky, 2009); cultural enclaves can often be found much closer to home. More generally, creative ideas may be stimulated if you strategically seek out cognitively challenging experiences. Try to learn a new language; spend time regularly with people whose norms and values and life experiences differ substantially from your own; visit with religious or political groups that are new to you. It can pay off.

A fourth rule of thumb: *Do things that you actually want to do*. This key to creativity is summed up nicely by Csikszentmihalyi (1997, p. S8): "Creative persons differ from one another in a variety of ways, but in one respect they are unanimous: They all love what they do." People are more creative when they do things that they find fun or enjoyable to do, and that they chose to do because of their personal interests or passions (Amabile, 1998). Best of all, people are more creative when they are happy (Baas et al., 2008). You are not just doing yourself a favor but may also be serving the broader goal of scientific innovation when you do things that you want to do. If you favor experiments, plan them. If you prefer applied work, apply yourself. If you prefer complex multivariate non-experimental analyses, disentangle away.

There will be times when you are unexcited, unhappy, and uninspired. Frustration, rejection, and bouts of burnout are common and normal, and there are good resources that provide advice on handling it (e.g., Jeremka et al., 2020). And you will sometimes be compelled to do things that other people think you *should* do rather than what you really *want* to do. Still, you can cultivate a creative mindset more effectively if you deliberately devote *some* of your time to activities that you are intrinsically motivated to do and that make you happy. After all, ideas are everywhere – in great books and trashy novels, television and movies, the lyrics of your favorite songs – and inspiration can strike not only when you're pouring over scientific papers but also when you're surfing the internet or walking in the woods or dancing with your friends. Some of these enjoyable activities might even be research projects. Designing a scientific study can be a fun. Designing a scientific study with your friends can be *really* fun. If you seek out projects that excite you, collaborators that you enjoy, and working environments that make you happy, you're more likely to be inspired with more ideas.

Idea-Generating Heuristics

Even if your mind is open, inspiration can be elusive. There are systematic strategies that scientists can use to develop worthwhile research ideas. McGuire (1997) provides a kind of catalog of strategies, identifying 49 “heuristics” that can be taught, learned, and used for the purpose of generating ideas. Many of these idea-generating heuristics involve reading the scientific literature and thinking systematically about what is and isn’t known. McGuire (1997) applied different labels to these different heuristics (e.g., “Reversing the Plausible Direction of Causality”; “Conjecturing Interaction Variables That Qualify a Relation”; “Generating Multiple Explanations for a Given Relation”). Fancy labels aside, these heuristics generally represent different ways of reacting thoughtfully to research results that seem to be not quite completely true – different ways of saying “Yes, but . . .”: Yes, research shows that *X* influences *Y*, but maybe *Y* influences variable *X* too? Research shows that *X* influences *Y*, but what if it sometimes doesn’t (i.e., the effect occurs only under some conditions or is limited to specific populations)? Research shows that *X* influences *Y*, but why (i.e., what is the underlying process? Is the proffered explanation the only plausible one?)? Research ideas can be generated by addressing such questions thoughtfully.

McGuire (1997) also identifies idea-generating heuristics that do *not* require reading scholarly literature, but instead involve attention to everyday life (e.g., “Recognizing and Accounting for the Oddity of Occurrences,” “Introspective Self-Analysis,” and “Sustained, Deliberate Observation”). There is an important principle underlying these heuristics: The goal of social and behavioral science research is to learn about the full scope of human behavior, and the scholarly literature is inevitably more narrow than that. Within the psychological sciences, for example, Berscheid (1992) describes how the important topic of close relationships was mostly ignored when most psychological scientists were men. Regardless of why these omissions exist, *they do*. It is limiting to look for ideas only within the scholarly literature. As a psychologist Nisbett (1990, p. 1078) wrote: “All of life is a source of psychological ideas” – but it’s an important principle that applies to all the social and behavioral sciences.

The important implication is that you can discover many fruitful ideas by raising your gaze from your scientific studies and casting it upon the real world instead. Cialdini (1980, p. 22) describes what happened when he took a break from puzzling over a frustratingly small effect observed on a rating scale and went to a football game:

> The crowd was suddenly up and shouting, and yelling encouragement to their favorites below. Arcs of tissue paper crossed overhead. The university fight song was being sung. A large group of fans repeatedly roared “We’re number one!” while thrusting index fingers upward. I recall quite clearly looking up from thoughts of that additional half unit of movement on a 7-point scale and realizing the power of the tumult around me. “Cialdini,” I said to myself, “I think you’re studying the *wrong* thing.”

The “*wrong* thing” was whatever that seven-point scale was failing to find. The *right* thing – the idea inspired by his fortuitous foray into the football stadium – turned into a productive multi-year program of research on group identification, self-esteem, and “basking in reflected glory.” Cialdini also made additional, more strategic observational journeys beyond the narrow halls of academe, such as the sabbatical he spent learning the tactics used by car dealers and pyramid scammers and other people whose real-life livelihoods depend on successful persuasion (Cialdini, 2006). These observations led to many new research projects and seminal contributions to the social and behavioral sciences.

A fifth rule of thumb. *Inspiration is idiosyncratic*. Some heuristics might work better for some people and others for others. Try everything and anything and remember Feyerabend's (1975): "*anything goes*."

Other People Are an Essential Source of Inspiration

There is a theme lurking in this chapter, and it merits being made explicit. Scientific research is a highly collaborative process, and most successful scientists operate within social networks of fellow scientists from whom they receive – and to whom they provide – social support (Perry-Smith & Mannucci, 2015). Other people are not only an asset when carrying out research projects, they are a great source of inspiration and ideas.

The sixth rule of thumb is perhaps the most important: *Interact with other people*. All of life is a source of ideas, and its corollary is that the more interesting the people you spend time with, the more interesting ideas you are likely to encounter (Nisbett, 1990). Close connections with other people serve as a catalyst for the generation of creative ideas, especially when those other people have diverse arrays of knowledge (Sosa, 2011). If you can forge those relationships within the context of the research that you do, it can make the research that you do feel less like work and a lot more fun. Rather than racking your brain in isolation in search of lonely inspiration, it might be more fun – and productive – to brainstorm research ideas with collaborators. The "catchy title" project started out mediocre, but conversations in the lab made the idea mature, catch fire, and become worthwhile.

Not your cup of tea to try McGuire's (1997) heuristics on your own? Try it over a cup of tea with a couple of friends. From modest beginnings, good ideas can grow. Science is a conversation; seek out opportunities to join it. Ask questions. Attend conferences. Talk with the people around you – students, teachers, friends, lovers, and maybe even strangers. If you can, find ways to ensure that the people around you have diverse interests, diverse attitudes, and diverse backgrounds. If you want to be graced by good ideas about how people feel, think, and live their lives (and by ideas about how to make their lives better), it helps to be actively engaged in people's lives. It helps to be a truly *social* scientist.

Strategies for Figuring out Whether an Idea is Worth Working on

You have an idea. Now what? A few pages ago we justified an "anything goes" attitude toward getting ideas with the observation that there would be time later to assess whether those ideas are any good or not. That time has come.

This kind of assessment is important. Research projects require a substantial investment of time – almost always more than you anticipate. The "catchy title" project began with one simple study and blossomed into a dozen more, some of which took months to complete – and that was a successful project that produced publishable results. Many research projects are less successful, but they still consume researchers' time and effort before they are abandoned. It is best to think carefully about whether an idea is worth pursuing before you do so.

How do you know which idea to pursue? That exact question was posed to some very productive psychological scientists some years ago (Dialogue, 2002). Their responses suggest that a wise decision about whether to pursue an idea (or not) is informed by answers to three important questions: (1) Is it interesting to you? (2) Is it interesting to other people? (3) Can you get it done?

Is It Interesting to You?

If you decide to pursue a research idea with an actual research project, you will devote a lot of your time and effort to that project. You will immerse yourself deeply in a scientific literature written with jargon and complexity. You will do the painstaking labor of designing a methodology, collecting data, and analyzing those data; ideally you will also do the painstaking labor of writing up the results in a manuscript and shepherding that manuscript into publication. Rarely does it all proceed as straightforwardly as you hope it will. Manipulations and measures may need to be pilot-tested, even multiple times. Before a manuscript is published, it may be rejected, often multiple times. Unless you have a will of steel and a disdain for reinforcement, your project is unlikely to succeed unless you are intrinsically motivated to see it through. There may be rewards along the way as well (e.g., new insights, new inspirations and ideas, the joys of surmounting a methodological challenge, learning a new data analytic technique, or making a novel scientific contribution), and these rewards too are more likely to accrue if you are truly passionate about the project. For all these reasons, this is a good place to repeat – and repurpose – one of the rules of thumb identified above: *Do things that you actually want to do*.

Successful scientists typically prioritize ideas that excite them personally. In that compendium of psychologists' responses to the question "How do you know which idea to pursue?" Brenda Major replied, "Does the idea grab me? Is it interesting? Can I get enthusiastic about it?"; and Elliot Aronson said, "I try to follow my own curiosity . . . to ask a researchable question that I am passionately interested in finding answers to" (Dialogue, 2002, p. 12). Some of these scientists advocated strategies to help assess whether initial interest might actually endure. Yoshihisa Kashima likes to imagine a future in which the initial idea has panned out perfectly – "hypotheses (or hunches) are supported, and everything is beautiful" – and then asks himself "Am I excited?" (p. 13). Anthony Greenwald offered the following pragmatic advice: "When you have a new research idea, try writing the title and abstract of the article that will report it. If (a) you can't write them or (b) you can write them but don't find them compelling, then abandon before you start" (p. 12). This kind of exercise can help you think about an idea more deeply – to consider it from multiple angles, to identify connections to existing lines of research, and perhaps even to generate additional ideas too. Interesting ideas often become even more interesting as you think about them more and more. If this doesn't happen for you, then perhaps it is not the idea for you.

Individuals' interests are idiosyncratic (we can expand that fifth rule of thumb: *both inspiration and interest are idiosyncratic*) and there are many reasons why you might be passionate about an idea. It doesn't matter why an idea excites you; what matters is that it does.

Is it Interesting to Other People?

It is a promising sign if an idea excites you, but that's only the beginning. It's important to ask whether an idea is interesting to other scientists and to people in general. There are both philosophical and practical reasons to ask this question.

Scientists don't do science in isolation. Philosophers define science not simply as an intellectual endeavor but as a fundamentally *social* activity involving a large number of people who, collectively, engage in inspection, criticism, disagreement, and discussion – that ultimately leads to progress (Grene, 1985; Longino, 1990; Thagard, 1978). Individual research projects are actually community projects; even a small research project is typically conducted by multiple people working in collaboration, using methodological strategies developed and refined by many other people, with the direct support of broader research communities (e.g., universities, funding agencies) and the indirect support of even larger communities (e.g., taxpayers, people who pay tuition). Scientists who draw upon those community resources have a responsibility to consider more than their own personal curiosity – they must also consider the interests of everyone else.

This philosophical perspective is complemented by purely a pragmatic consideration. Regardless of results, and regardless of your personal interest, your research project is unlikely to be published (or to make any kind of meaningful contribution) if that research is of interest only to you. The underlying ideas must interest other people too.

Some topics are more generally interesting than others. Topics such as altruism, depression, language acquisition, religious belief, and social status have been of broad and enduring interest, whereas other topics may be more faddish or of interest only to niche audiences. To some extent, these differences reflect differences in conceptual scope and range of applicability (van Lange, 2013). Scientists' interests also reflect real-world relevance. Although some social and behavioral scientists – especially psychologists – use contrived methods in controlled laboratory environments, the phenomena under inquiry are expected to reflect the real world. The more this connection is evident, the more other scientists (and non-scientists) are likely to find an idea interesting. Some research ideas have transparent implications for useful real-life applications, including applications that might help to solve social problems, promote health and well-being, or to otherwise improve humans' lives. People are likely to find these kinds of ideas important and, therefore, interesting. Cialdini observed "if there is evidence that the effect occurs regularly and powerfully in multiple environments, it is simply more worthy of examination." Similarly, Aronson said "From time to time, as a researcher, I ask myself: "Is this research ever going to do anyone any good?" (Dialogue, 2002, p. 13). These observations lead us to a seventh rule of thumb: *If an idea has more real-life relevance, people are more likely to be interested in it*.

And before you can catch your breath, we offer an eighth rule of thumb: *If an idea is too obviously true, people might not find it interesting*. Because scientists value veracity, it might be tempting to think that the more obviously true some hunch or hypothesis is, the more obviously interesting it will be; that's not the case. Davis (1971) argued that the subjective experience of surprise is a critical component of subjective interest value and that people are more likely to consider a scientific proposition to be interesting if it challenges some presumption that they have previously taken for granted. According to Davis (1971, p. 313), the essential formula for an interesting idea can be expressed semi-algebraically: "What seems to be *X* is in reality non-*X*" or "What is accepted as *X* is actually non-*X*." A good example of this is the discovery that partial reinforcement leads to more durable performance than continuous reinforcement – less *is* more (Skinner, 2019).

This principle helps to explain scientists' attraction to counter-intuitive ideas (Gray & Wegner, 2013). In fact, researchers in some social and behavioral science fields have been criticized for being a little *too* fond of counter-intuitive phenomena – and for not attending closely enough to the real possibility that results that violate conventional presumptions of truth might actually be false (Yong, 2012). But the most interesting and useful ideas are not merely counter-intuitive; they provide a way to resolve the apparent conflict between an existing presumption (X) and a challenging new proposition (non-X). Galen Bodenhausen observed "Interesting ideas often have elements that are surprising and, at least at the first pass, difficult to reconcile with one's most immediately relevant knowledge structures, but in bringing other knowledge to bear in a novel way, the inconsistencies are resolved in a way that can have an intellectually satisfying elegance . . . that marks an idea as interesting and worthy of pursuit" (Dialogue, 2002, pp. 12–13).

Ideas do not need to be counter-intuitive to fit Davis's (1971) formula. For instance, the results of replication studies are rarely considered to be counter-intuitive, but the ideas underlying replication research can still fit that formula. A phenomenon presumed to be robust and replicable may not be so robust or easy to replicate after all. A phenomenon presumed to be of questionable replicability may be revealed to be replicable after all (e.g., Noah et al., 2018). There are many ways in which ideas may challenge people's preconceptions. Savvy scientists think carefully about what those preconceptions are and about whether and how an idea might challenge them.

Some ideas may be so unconventional that they might seem implausible or even incomprehensible, and that too is a barrier to attracting others' interest. The most successful ideas are often those that occupy the sweet spot between the extremes of obvious and outlandish. Marilynn Brewer characterized this sweet spot as a kind of *optimal distinctiveness*: "does the idea seem grounded in current research (i.e., have a degree of familiarity) and yet hasn't already been introduced in the recent literature (i.e., have a degree of novelty)" (Dialogue, 2002, p. 14). Daniel Gilbert also highlights this sweet spot, while also neatly summarizing a handful of other characteristics that make ideas interesting to other people (Dialogue, 2002, p. 14):

> A good idea is original, tractable, economical, synthetic, generative, and grand. By that I mean it is not well-explored (original), it is explorable with scientific methods (tractable), it provides an elegant and simple solution to a complex set of problems (economical), it brings together phenomena that initially seemed to have nothing in common (synthetic), it generates many more interesting questions than it answers (generative), and it speaks about some fundamental truth (grand). Good ideas are almost never outlandish: When someone tells you a really good idea, you almost always have the sense that you were just about to think of it yourself except that . . . well, you didn't.

Any idea can be scrutinized for interest, and this process benefits from familiarity with relevant scholarly literatures. A thorough reading of those literatures? Daunting. You must do the deep dive eventually (if you actually do pursue the idea), but it is rarely the best place to begin. A more efficient way to begin is to bounce the idea off other people. Science is a conversation, and a potentially promising idea is a great conversation-starter. Talk about the idea with experts; even established scholars are usually happy to discuss ideas, especially if you are well-prepared and succinct. Talk about the idea with people who *aren't* experts. Their perspectives – along with their questions, criticisms, and occasional confusions – will help focus the idea, sharpen it, and clarify exactly what it is and why it matters. Nisbett (1990, p. 1082) made this plain: "The necessity of explaining one's concerns to others, and of putting them into a broader context, together with the effort to demonstrate why certain topics are interesting, all have the most direct benefits for thinking about research." The benefits are many. If an idea withstands public scrutiny and remains interesting, it may be worth pursuing.

These conversations can help refine the idea, reveal non-obvious nuances that make it more interesting, or identify important real-life applications that might make it even more worthwhile to pursue. If you can excite other intelligent people with an idea – and maybe recruit them as research collaborators – the resulting research project is likely to be more fun *and* successful.

Can You Get It Done?

You have a research idea that excites you and others. You are confident that the research – if done rigorously and well – will make a worthwhile scientific contribution. Someone should do it. Should that someone be *you*?

Before starting any research project, it is sensible to think about it from a purely pragmatic perspective – to consider not only the rewards it might bring to you (e.g., pleasure, publications) but also the resources required to pull it off. Some research projects are cheap to do. Others are not and may require extraordinary resources – special personnel, expensive equipment, dedicated laboratory space, access to exceptional populations, that sort of thing. Can you realistically acquire these resources? Do you have colleagues with connections? Can you write a grant application with a reasonably high probability of success? Can you do so in a timely way?

Time is a cost that you would be wise to consider carefully (and not just because people who place a high value on time are happier than those who don't; Whillans et al., 2016). The time spent on any research project is time that cannot be spent on anything else that might matter to you, including other potentially rewarding research projects. Regardless of the number of hours you personally spend on a project, some projects take [much] longer to complete than others. This can be an important consideration, perhaps especially important depending on your circumstances. Tenured professors may have the luxury of pursuing a project that might take years to pay off; untenured faculty and graduate students might not. When Chris Crandall was in graduate school, he chose – perhaps optimistically – to pursue a longitudinal field study for his dissertation. It took three years to complete and delayed (by a year) the completion of his PhD. It paid off, but plenty of equally time-consuming projects don't.

You would be wise to consider these kinds of costs carefully and to consult with other people about them. If, after doing so, you are convinced that you are the right person to pursue a research idea, go ahead and do it. If not, you might want to pursue a less costly project instead. That doesn't mean that you should just abandon entirely the costlier idea. Perhaps you will have the opportunity to return to it sometime in the future when you can more readily afford the costs. Some good ideas can wait, but don't trust your memory (*write it down*).

Science is a community project, but individual human beings are the vessels through which scientific ideas and empirical results must travel. Any decision about whether an idea is worth working on (or not) is a personal decision that will be informed by your own idiosyncratic interests, constraints, and aspirations. With that in mind, we give the last word here to the editors who solicited, and compiled, successful psychologists' thoughts about these decisions (Dialogue, 2002, p. 15):

> Which idea to pursue must depend upon your own goals . . .. If you want to publish a large number of articles in a reasonable amount of time, then one might pursue moderately novel ideas. If you want to have a lot of impact, then pursue innovative and contrarian ideas in a currently hot topic. If you want a grant, then focus on ideas that will pay off in a straightforward way in a reasonable amount of time (and money). If you want to enjoy your work, then follow your heart. These are not necessarily mutually exclusive.

Strategies for Transforming an Idea into Something Scientific

You've got an idea and you're excited to pursue it. The idea is taking shape not only in the form of an interesting research question but maybe also a speculative answer – your hunch about how the world works or your personal prediction about some relation between some set of variables. You might even be talking about your "hypothesis." Not so fast! There is work to be done. No matter how compelling your idea, no matter how convinced you are that your hunch or personal prediction might be right, it may not yet rise to the level of rigor that characterizes good science.

Scientific inquiry is characterized by methodological rigor – by methods that are systematic and precise and that are designed to minimize the impact that scientists' biases, blind spots, and subjective beliefs might have on scientific knowledge. People are accustomed to applying these principles to the *empirical* part of the scientific process, during which scientists collect and analyze data to test scientific conjectures. Less obviously, the same principles can be applied to the *conceptual* part of the process – the part in which scientists develop and articulate those conjectures in the first place. Among the many elements that characterize scientific rigor (e.g., Casadevall & Fang, 2016), there are two elements that you might be especially mindful of when developing a research idea into something that meets the high standards of science: *precision* and *impartiality*. These can transform a vague idea into a good idea.

Precision

"To ask a scientific question about individual or social behavior, we must specify the parts of a system and the relationships between them . . . The precise specification of parts and relationships is what defines a scientific question and separates it from wishy-washy pseudotheory" (Smaldino, 2017, pp. 314–315). Precise specification is a non-trivial challenge in the social and behavioral sciences because the "parts" of conceptual interest – *constructs* such as resilience, social status, or moral reasoning – are broad in scope and abstract in principle. They tend to be understood intuitively but imprecisely. For example, one person's intuitive understanding of "resilience" may only approximately match someone else's understanding of it. Unless these constructs are defined transparently and precisely, problems may arise in the form of mismatches between the empirical methods people use and the constructs of actual interest. To test an idea about "social status," you might sensibly use a measure that someone else had used to measure social status without realizing that it measures something different from the sort of social status that you had in mind. Two people may attempt to test the same hypothesis about social status but have different intuitive understandings of social status and consequently use different measures that produce different results – creating the superficial appearance of inconsistent support for a conceptual hypothesis when, in reality, that hypothesis might actually have only been meaningfully tested by one (or none) of the studies (Oberauer & Lewandowsky, 2019). Your initial ideas and hunches are unlikely to be characterized by the level of conceptual precision required to avoid these problems.

The goals of transparency and precision lead us to a ninth rule of thumb: *Before pursuing any idea seriously, precisely define its conceptual "parts," and make clear their relations to each other.* You – and anyone who reads or listens to you – should be able to articulate clearly what each relevant construct is and is not.

Formal modeling methods can help with this task (Smaldino, 2017). Also helpful are systematic methods of construct validation (Clark & Watson, 2019; Grahek et al., 2021). It is tempting to think that the proper time to consider construct validity is only after an idea has been formulated and a scientist has begun designing an empirical study. *This is wrong*. A precise conceptual definition of a construct is necessary right from the get-go. Clark and Watson (2019, p. 1413) wrote "an essential early step is to crystallize one's conceptual model by writing a precise, reasonably detailed description of the target construct"; and they provide useful guidance. Try asking a few simple questions about every construct you work with. What exactly is it? What isn't it? In what specific ways does this construct overlap with and differ from other similar constructs? Is this construct truly a single coherent thing or are there different varieties that deserve their own distinct conceptual definitions (and empirical operationalizations)? This kind of systematic conceptual work takes careful thought and effort, but, as Grahek et al. (2021, p. 811) observe, "the effort can pay off in the form of more precise conceptual definitions of constructs (and, consequently, better measures of those constructs), more carefully articulated theories about those constructs, and more nuanced hypotheses that make accurate predictions."

Impartiality

People sometimes think that a scientific hypothesis is much the same thing as a scientist's own personal prediction. Philosophers of science beg to differ. Karl Popper (1959/2005) made a sharp distinction between a truly scientific conjecture (e.g., an objective statement stipulating some logically plausible relation between constructs) and scientists' subjective beliefs about whether that conjecture is true or not. A hunch or personal prediction is indistinguishable from a subjective belief, and simply calling it a "hypothesis" does not make it so. It typically takes careful logical analysis to transform an informal idea or personal prediction into a rigorously objective scientific hypothesis.

In addition to high standards of scientific rigor, there is also a purely pragmatic reason to engage in this kind of systematic logical analysis. It can help you make well-informed decisions when designing studies to test hypotheses – increasing the likelihood that these studies will produce useful data, replicable results, accurate inferences, and publishable papers.

When people perceive something to be their own personal creation or personal possession, they overestimate its value (e.g., Morewedge et al., 2009). The implication is that when people personalize a hypothesis ("my hypothesis"), they are more likely to believe that it's true even if it's not. In addition, if the hypothesis is true, they are more likely to overestimate the size of the effect and the extent to which it generalizes across different circumstances or populations. These kinds of overestimates can lead researchers to make problematic decisions when designing studies and analyzing data (Schaller, 2016, p. 109):

> When researchers overestimate the veracity of hypothesized effects, they are less likely to make the kind of decisions (in data analytic strategies and subsequent reporting of empirical results) that guard against the documentation of false-positive inference. When researchers overestimate the size of hypothesized effects, they are more likely to employ underpowered research designs – increasing the likelihood

that, whenever effects are detected, they are likely to be erroneously big. And when researchers overestimate the generalizability of hypothesized effects, they are less likely to empirically test its context-specificity or to otherwise draw attention to its potential fragility.

To avoid falling prey to these problems, it helps to adopt an impartial attitude toward ideas, predictions, and hypotheses. Can you be impartial even when doing research on topics of great personal interest to you? Yes! You can be personally interested in a research *question* while still cultivating an impartial attitude regarding the accuracy of hypothetical *answers* to that question. As a scientist, passionate interest in an idea need not – and should not – supersede your passion for honesty, accuracy, and truth. If you cannot accept reliable findings, you'll need to examine your commitments.

Let us revisit that lovely line from García Márquez (1995, p. 56) and reframe it as the tenth rule of thumb: *Ideas do not belong to anyone*. You may be a more effective steward of ideas and hypotheses – and make wiser decisions when testing them – if you adopt the mindset that you are steward (and not owner) of those ideas. You may have your informal hunches and subjective beliefs, but they are distinguishable from scientific hypotheses. To be scientific hypotheses, conjectures must be stated impartially. To be *compelling* hypotheses, they require careful and coherent justification.

A useful pathway to transforming an informal idea into an impartially stated, carefully justified scientific hypothesis leads us to one last rule of thumb: *Specify your assumptions explicitly*. Try to identify all the assumptions underlying a personal prediction and then derive a clearly stated and testable hypothesis from these assumptions, using a sequence of "if–then" statements (Schaller (2016) offers examples). If you cannot get a hypothesis to follow logically from the assumptions, it might be a clue that your hunch is wrong or perhaps you haven't yet specified precisely *why* it might be right. Have you failed to specify a key assumption? Is there a necessary logical step that you intuitively appreciate but haven't yet articulated? Connecting those logical dots makes a more convincing case that the hypothesis is not merely an idiosyncratic hunch but is a plausible scientific hypothesis.

This explicit identification and systematic inspection of underlying assumptions and derivations help forecast the plausibility, size, and generalizability of hypothesized effects. This leads one to make better choices for empirical research (e.g., sample sizes, measurement strategies, and power of manipulations). Is every assumption and logical derivation completely convincing? If not, this is a reminder to maintain skepticism (a key scientific value) toward the hypothesis you're developing and guard against confirmation bias. Does each assumption and if–then statement apply equally to *all* people under *all* circumstances? If not, the overall hypothesis may accurately describe some people but not others or may be true under some circumstances but not others. This information too can inform methodological decision-making (e.g., decisions about specific populations to sample or about specific moderating variables to manipulate or measure) and may lead you to new ideas and new, more nuanced (i.e., better) scientific hypotheses. That's been your goal all along.

Does every scientist subject their ideas to this kind of systematic logical analysis? Alas, no. Compared to our scholarly cousins in the physical, biological, and cognitive sciences, many social and behavioral scientists have tended to be looser and lazier about articulating hypotheses with precision and rigor, *but that's changing*. Scientists are increasingly aware of the problems that arise from informal conceptual analyses and the benefits that accrue from the extra work required to transform inspiration and intuition into precise, carefully articulated, and logically transparent statements that meet high standards of scientific rigor (e.g., Fiedler, 2107; Gervais, 2021; Grahek et al., 2021; Gray, 2017; Klein, 2014; Muthukrishna & Henrich, 2019; Oberauer & Lewandowsky, 2019; Schaller, 2016; Smaldino, 2017, 2020). It's an important part of the ongoing effort to do science better.

You might reasonably ask: Isn't all this extra effort time-consuming (and sometimes tedious) to do? Yes – and that's a clue that it can be good to do. Compared to less rigorous means of inquiry, more rigorous methodological practices are, inevitably, more time-consuming (and sometimes tedious). That's science.

But you don't have to do it all by yourself, and it's best if you don't. The kind of painstaking conceptual work that we have described here (i.e., precise definitions of abstract constructs, detailed logical dissections of hypotheses) is likely to be more productive – and more fun – if you do it in collaboration with other people. It's a good way to do good science.

Envoi

Scientists love new research ideas, and so it is ironic that scientists receive so little formal education about how to find new ideas and develop them rigorously. To the extent that scientists get this guidance, it is haphazard and idiosyncratic ("the apprenticeship model"), consisting of informal discussions with mentors and peers, feedback (sometimes fulsome and constructive, often not) from reviewers, brevity-is-the-soul-of-wit editors, committees, and a lot of reading between the lines. A few books and articles provide useful guidance of one sort or another (e.g., Beveridge, 1957; McGuire, 1997; Nisbett, 1990), and young scholars can learn a lot from anecdotes and personal reflections that are sometimes compiled in out-of-the-way places (e.g., Dialogue, 2002).

We have drawn upon these and other sources (such as the psychological research on creativity and philosophy of science) to identify strategies – and guiding principles – that might be helpful. Getting ideas and making the most of them takes more than idle inspiration – they benefit from strategy, skill, and labor. Science – as practice, as a profession, as a cultural product – does not usually come easily. Still, most people are well equipped to meet those challenges. Curiosity is natural. Opportunities for inspiration are everywhere. The skill set required to transform informal ideas into useful scientific products is attainable. Most of these challenges can be more readily surmounted by using one simple trick: *talk to other people*. Time, training, practice, and talk make the "idea" part of science easier, and you get better at it.

We close with a snippet of conversation from two people who are very good at it: Shelley Taylor and Susan Fiske (Taylor & Fiske, 2019, p. 8):

Susan Fiske: Do you have any suggestions for people starting out in the field about how to have a good idea, and how to implement it?

Shelley Taylor: I have always thought that you look around you and if you're psychologically minded, you notice things, and you think, *Well, what does that mean?* You keep trying to step it up a level, which will ultimately lead you to theory. I would say trusting your own ideas is a very important way of coming up with a research program that is novel and exciting and that ultimately wins people over.

Susan Fiske: I think that's a great place to end.

References

Amabile, T. M. (1998). How to kill creativity. *Harvard Business Review*, *76*, 76–87.

Baas, M., De Dreu, C. K. W., & Nijstad, B. A. (2008). A meta-analysis of 25 years of mood–creativity research: Hedonic tone, activation, or regulatory focus? *Psychological Bulletin*, *134*, 779–806. https://doi.org/10.1037/a0012815

Bahns, A. J., Crandall, C. S., Gillath, O., & Preacher, K. J. (2017). Similarity in relationships as niche construction: Choice, stability, and influence within dyads in a free choice environment. *Journal of Personality and Social Psychology*, *112*(2), 329–355.

Berscheid, E. (1992). A glance back at a quarter century of social psychology. *Journal of Personality and Social Psychology*, *63*, 525–533. https://doi.org/10.1037/0022-3514.63.4.525

Beveridge, W. I. B. (1957). *The Art of Scientific Investigation*. Vintage Books.

Campbell, D. T. (1974). Evolutionary epistemology. In P. A. Schilpp (ed.), *The Philosophy of Karl Popper* (vol. 1, pp. 413–463). Open Court.

Casadevall, A. & Fang, F. C. (2016). Rigorous science: A how-to guide. *mBio*, *7*, e01902–e01916. https://doi.org/10.1128/MBIO.01902-16

Cialdini, R. B. (1980). Full-cycle social psychology. In L. Bickman (ed.), *Applied Social Psychology Annual, Volume 1* (pp. 21–47). SAGE Publications.

Cialdini, R. B. (2006). *Influence: The Psychology of Persuasion*. Harper Business.

Clark, L. A. & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, *31*, 1412–1427. https://doi.org/10.1037/pas0000626

Csikszentmihalyi, M. (1997). Happiness and creativity. *The Futurist*, *31*, S8–S12.

Davis, M. S. (1971). That's interesting! Towards a phenomenology of sociology and a sociology of phenomenology. *Philosophy of the Social Sciences*, *1*(2), 309–344.

Dialogue (2002). Which scientific problem to pursue? Eminent social/personality psychologists reveal their secrets of scientific success to the editors of Dialogue. *Dialogue*, *17*(2), 12–15. https://spsp.org/sites/default/files/dialogue172.pdf

Feist, G. J. (1998). A meta-analysis of the impact of personality on scientific and artistic creativity. *Personality and Social Psychology Review*, *2*, 290–309. https://doi.org/10.1207/s15327957pspr0204_5

Feyerabend, P. (1975). *Against Method*. New Left.

Fiedler, K. (2017). What constitutes strong psychological science? The (neglected) role of diagnosticity and a priori theorizing. *Perspectives on Psychological Science*, *12*, 46–61. https://doi.org/10.1177/1745691616654458

García Márquez, G. (1995). *Of Love and Other Demons*. Knopf. (Originally published as *Del amor y otros demonios* by Mondadori. Translated by Edith Grossman.)

Gervais, W. M. (2021). Practical methodological reform needs good theory. *Perspectives on Psychological Science*, *16*, 827–843. https://doi.org/10.1177/1745691620977471

Grahek, I., Schaller, M., & Tackett, J. L. (2021). Anatomy of a psychological theory: Integrating construct validation and computational modeling methods to advance theorizing. *Perspectives on Psychological Science*, *16*, 803–815. https://doi.org/10.1177/1745691620966794

Gray, K. (2017). How to map theory: Reliable methods are fruitless without rigorous theory. *Perspectives on Psychological Science*, *12*, 731–741. https://doi.org/10.1177/1745691617691949

Gray, K. & Wegner, D. M. (2013). Six guidelines for interesting research. *Perspectives on Psychological Science*, *8*, 549–553. https://doi.org/10.1177/1745691613497967

Grene, M. (1985). Perception, interpretation, and the sciences. In D. J. Depew & B. H. Weber (eds.), *Evolution at a Crossroads: The New Biology and the New Philosophy of Science*. MIT Press.

Hull, D. L. (1988). *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science*. University of Chicago Press.

Jaremka, L. M., Ackerman, J. M., Gawronski, B., et al. (2020). Common academic experiences no one talks about: Repeated rejection, impostor syndrome, and burnout. *Perspectives on Psychological Science*, *15*, 519–543. https://doi.org/10.1177/1745691619898848

Klein, S. B. (2014). What can recent replication failures tell us about the theoretical commitments of psychology? *Theory & Psychology*, *24*, 326–338. https://doi.org/10.1177/0959354314529616

Lam, T. W. H. & Chiu, C.-Y. (2002). The motivational function of regulatory focus in creativity. *Journal of Creative Behavior*, *36*, 138–150. https://doi.org/10.1002/j.2162-6057.2002.tb01061.x

Leung, A. K.-Y., Maddux, W. W., Galinsky, A. D., & Chiu, C.-Y. (2008). Multicultural experience enhances creativity: The when and how. *American Psychologist*, *63*, 169–181. https://doi.org/10.1037/0003-066X.63.3.169

Longino, H. E. (1990). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press.

Maddux, W. W. & Galinsky, A. D. (2009). Cultural borders and mental barriers: The relationship between living abroad and creativity. *Journal of Personality and Social Psychology*, *96*, 1047–1061. https://doi.org/10.1037/a0014861

McGuire, W. J. (1997). Creative hypothesis generating in psychology: Some useful heuristics. *Annual Review of Psychology*, *48*, 1–30. https://doi.org/10.1146/annurev.psych.48.1.1

Morewedge, C. K., Shu, L. L., Gilbert, D. T., & Wilson, T. D. (2009). Bad riddance or good rubbish: Ownership and not loss aversion cause the endowment effect. *Journal of Experimental Social Psychology*, *45*, 947–951. https://doi.org/10.1016/j.jesp.2009.05.014

Murray, D. R. & Schaller, M. (2016). The behavioral immune system: Implications for social cognition, social interaction, and social influence. *Advances in Experimental Social Psychology*, *53*, 75–129. https://doi.org/10.1016/bs.aesp.2015.09.002

Muthukrishna, M. & Henrich, J. (2019). A problem in theory. *Nature Human Behavior*, *3*, 221–229. https://doi.org/10.1038/s41562-018-0522-1

Niiniluoto, I. (2019). Scientific progress. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2019 Edition). Available at: https://plato.stanford.edu/archives/win2019/entries/scientific-progress/.

Nisbett, R. E. (1990). The anti-creativity letters: Advice from a senior tempter to a junior tempter. *American Psychologist*, *45*, 1078–1082.

Noah, T., Schul, Y., & Mayo, R. (2018). When both the original study and its failed replication are correct: Feeling observed eliminates the facial-feedback effect. *Journal of Personality and Social Psychology*, 114(5), 657–664. https://doi.org/10.1037/pspa0000121

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, *26*, 1596–1618. https://doi.org/10.3758/s13423-019-01645-2

Perry-Smith, J. & Mannucci, P. V. (2015). Social networks, creativity, and entrepreneurship. In C. E. Shalley, M. A. Hitt, & J. Zhou (eds.), *The Oxford Handbook of Creativity, Innovation, and Entrepreneurship* (pp. 205–224). Oxford University Press.

Popper, K. (1959/2005). *The Logic of Scientific Discovery*. Routledge. (Originally published in 1934 as *Logik der Forschung* by Verlag von Julius Springer.)

Popper, K. (1963). *Conjectures and Refutations*. Routledge and Keagan Paul.

Root-Bernstein, R. & Root-Bernstein, M. (2004). Artistic scientists and scientific artists: The link between polymathy and creativity. In R. J. Sternberg, E. L. Grigorenko, & J. L. Singer (eds.), *Creativity: From Potential to Realization* (pp. 127–151). American Psychological Association.

Schaller, M. (2016). The empirical benefits of conceptual rigor: Systematic articulation of conceptual hypotheses can reduce the risk of non-replicable results (and facilitate novel discoveries too). *Journal of Experimental Social Psychology*, *66*, 107–115. https://doi.org/10.1016/j.jesp.2015.09.006

Skinner, B. F. (2019). *The Behavior of Organisms: An Experimental Analysis*. Appleton-Century. (First published by Appleton-Century Company in 1938).

Smaldino, P. E. (2017). Models are stupid, and we need more of them. In R. Vallacher, S. Read, & A. Nowak (eds.), *Computational Social Psychology* (pp. 311–331). Routledge. https://doi.org/10.4324/9781315173726-14

Smaldino, P. E. (2020). How to translate a verbal theory into a formal model. *Social Psychology*, *51*, 207–218. https://doi.org/10.1027/1864-9335/a000425

Sosa, M. E. (2011). Where do creative interactions come from? The role of tie content and social networks. *Organization Science*, *22*(1), 1–21. https://doi.org/10.1287/orsc.1090.0519

Taylor, S. E. & Fiske, S. T. (2019) Interview with Shelley E. Taylor. *Annual Review of Psychology*, *70*, 1–8. https://doi.org/10.1146/annurev-psych-041818-040645

Thagard, P. R. (1978). Why astrology is pseudoscience, *PSA*, *1*, 223–234.

Van Lange, P. A. M. (2013). What we should expect from theories in social psychology: Truth, abstraction, progress, and applicability as standards (TAPAS). *Personality and Social Psychology Review*, *17*, 40–55. https://doi.org/10.1177/1088868312453088

Whillans, A. V., Weidman, A. C., & Dunn, E. W. (2016). Valuing time over money is associated with greater happiness. *Social Psychological and Personality Science*, *7*, 213–222. https://doi.org/10.1177/1948550615623842

Yong, E. (2012). Replication studies: Bad copy. *Nature*, *485*, 298–300.

4 Literature Review

Rachel Adams Goertel

Abstract

A literature review is a survey of scholarly sources that establishes familiarity with and an understanding of current research in a particular field. It includes a critical analysis of the relationship among different works, seeking a synthesis and an explanation of gaps, while relating findings to the project at hand. It also serves as a foundational aspect of a well-grounded thesis or dissertation, reveals gaps in a specific field, and establishes credibility and need for those applying for a grant. The enormous amount of textual information necessitates the development of tools to help researchers effectively and efficiently process huge amounts of data and quickly search, classify, and assess their relevance. This chapter presents an assessable guide to writing a comprehensive review of literature. It begins with a discussion of the purpose of the literature review and then presents steps to conduct an organized, relevant review.

Keywords: Literature Search, Conceptual Saturation, Systematic Review, Narrative Review, Integrative Review

> If I have seen a little further it is by standing on the shoulders of giants.
>
> Isaac Newton in a letter to his rival Robert Hooke, 1676

Introduction

This chapter seeks to demystify the complexity of the literature review by breaking down a notoriously monumental undertaking into manageable steps. Before conducting any research, relevant past research must be selected, analyzed, synthesized, and contextualized within existing knowledge. The purpose of the review is to determine what is known about the topic and disclose gaps that may exist in the research. Comprehensive analysis is a significant portion of this process. A literature review does not just summarize sources – it systematically analyzes, synthesizes, and critically evaluates seminal and current research to give a clear picture of the state of knowledge on the subject. An inclusive review can even take a meta-analytic approach by comparing the results of individual research studies to identify patterns and trends and/or detect sources of dissimilarity among similar studies (see Chapter 27 in this volume). It uses statistical methods to analyze and summarize research while a traditional review answers a defined research question by collecting and summarizing all empirical evidence that fits pre-specified eligibility criteria. Regardless of the approach, the research literature collected, organized, analyzed, and synthesized results in a manuscript called a review of literature or literature review.

Understanding the Literature Review

A literature review is based on an extensive critical examination and synthesis of the relevant literature on a topic and includes a critical analysis of the relationship among these different works, while relating findings to the project at hand. Therefore, it is a crucial aspect of research proposals for grants, dissertations, and theses and serves as the framework for research papers.

Purpose

The aim of a review is not only to assess the research but also to increase researchers' understanding of research during the review process. "All researchers explore the literature for material about their topic; first to see what has already been done and second to profit from findings, cautions and suggestions made by other researchers" (Mertler and Charles, 2011, p. 63). In a well-done literature review, a researcher shows familiarity and understanding with a body of knowledge about a topic and thereby establishes credibility. The literature review shows how previous research is linked to the project by summarizing and synthesizing what is known while identifying gaps in the knowledge base, facilitating theory development, closing areas where enough research already exists, and uncovering areas where more research is needed (Webster & Watson, 2002, p. xiii). The literature review gives the researchers a chance to:

- demonstrate familiarity with the topic and scholarly context of current and past research
- develop a theoretical framework and methodology for research
- provide an intellectual context for their own work
- position and evaluate themselves in relation to other researchers and theorists
- show how their research addresses a gap or contributes to a debate
- avoid redundancy by saving time researching something that has already been done.

In its most general form, a literature review is a survey of scholarly sources that provides an overview of a particular topic. It is an organized, synthesized collection of the most relevant and significant research regarding that topic to provide a comprehensive look at what has been said on the topic and by whom.

Application

Once complied, synthesized, and written, the review of literature goes well beyond an overview of the reviewed literature. It is indispensable in academic research. "A substantive, thorough, sophisticated literature review is a precondition and the foundation for doing substantive, thorough, sophisticated research. A researcher cannot perform significant research without first understanding the literature in the field" (Boote & Beile, 2005, p. 3). The purpose is not only to summarize but also to synthesize the arguments and ideas of others to support a new insight that the literature reviewer will contribute to the field. According to Cooper (1984), the worth of any single study is derived as much from how it fits with and expands on prior research as from the study's intrinsic contributions

A literature review may consist of a summary of key research sources, but typically in the social and behavioral sciences a literature review has an organizational pattern that presents a synthesis within specific conceptual categories. The literature review can be used for a range of purposes, including detection of gaps, identification of areas for further study, or guidance for evidence-based practice. Reviewers critically examine the literature either as separate projects related to no other purpose other than to review the current research or as standing as a part of a paper that aims to make a contribution (Rhoades, 2011). Thus, reviews can be a valuable contribution to any research investigation. They may form the basis of developing standards and guidelines for practice as well as policies, procedures, and innovation in a particular field of study (McCabe, 2005, p. 41). They provide important insight into a particular scholarly topic and are considered an essential tool.

Types of Literature Reviews

There are many types of literature reviews including argumentative, integrative, methodological, etc. The social and behavioral sciences tend to have three main types: the systematic, the narrative, and the integrative review. Snyder (2019) outlines the three types, which are summarized below:

Systematic Review

The systematic review synthesizes research findings in an organized, transparent, and reproducible way. The aim of a systematic review is to identify evidence to address a particular research question. The methods of systematic reviews involve developing eligibility criteria and describing information sources, search strategies, study selection processes, outcomes, assessment of bias in individual studies, and data synthesis (Moher et al., 2015). In other words, a systematic review asks a specific question and tries to answer it by summarizing evidence that meets a set of pre-specified criteria. The process starts with a research question and a protocol or research plan. The researcher seeks relevant studies that are screened for eligibility using their inclusion and exclusion criteria. Next, the reviewer extracts the relevant data and assesses the quality of the included studies. Finally, the reviewer synthesizes the extracted study data and presents the results.

The systematic review generally:

- has clearly stated objectives with predefined eligibility criteria for studies
- uses explicit, reproducible methodology
- employs a systematic search that attempts to identify all studies
- assures an assessment of the validity of the findings of the included studies (e.g., risk of bias)
- provides a systematic presentation, and synthesis, of the characteristics and findings of the included studies.

Narrative Review

The narrative review aims for an overview of a topic and examines how research within a selected field has progressed over time. In general, the narrative review seeks to identify and to understand potentially relevant research and to synthesize this narrative research to provide clarity about complex areas. Narrative reviews are a discussion of topics from a theoretical point of view. These reviews take a less formal approach in that narrative reviews do not require the more rigorous aspects of a systematic review, such as reporting methodology, search terms, databases used, and inclusion and exclusion criteria (Nobre et al., 2003). This type of analysis can be useful for identifying themes, theoretical perspectives, or common issues within a specific research topic, which are considered important tools in continuing education and research. A narrative review could manifest in a historical overview, synthesize the state of knowledge, or map a field of research. It generally:

- starts with a clear question to be answered, but more often involves a general discussion of a subject with no stated hypothesis
- does not usually attempt to locate all relevant literature but rather utilizes pivotal papers
- employs subjectivity in study selection that potentially leads to biases.

Integrative Review

The integrative review, also referred to as the critical review, seeks to assess, critique, and synthesize the literature on a topic in a way that enables new theoretical frameworks or perspectives to emerge (Torraco, 2005, p. 358). An integrative review looks broadly at a phenomenon and allows for diverse research that may contain theoretical and methodological literature as well as both quantitative and qualitative studies. This approach supports a wide range of inquiry, such as defining concepts, reviewing theories, or analyzing methodological issues. Similar to the systematic review, it uses a systematic process to identify, analyze, appraise, and synthesize all selected studies, but does not include statistical synthesis methods. Furthermore, the aim of an integrative review is to generate a new conceptual framework or theory and focuses on the advancement of knowledge (p. 335). The integrative review generally:

- includes five distinct steps including (1) problem formulation, (2) data collection or literature search, (3) evaluation of data, (4) data analysis, and (5) interpretation and presentation of results
- maintains scientific integrity while conducting an integrative research review with careful consideration of threats to validity.

This classification scheme does not privilege any specific type of review as being more valued. As explained above, each type of review has its own strengths and limitations. The significance of a comprehensive literature review that synthesizes research findings is valuable. Many authors agree that especially systematic, stand-alone literature reviews can make an important contribution to existing research (Boote and Beile, 2005).

Other Types of Reviews

Of note, there are other types of literature reviews besides the systematic, narrative, and integrative. Those less common reviews include: meta-analysis – a systematic review that takes findings from several studies on the same subject and analyzes them using standardized statistical procedures (see Chapter 27 in this volume); scoping – an assessment of the potential scope of the research literature on a particular topic that helps determine gaps in the research; and conceptual – a group research effort to examine concepts or themes to identify the current understanding of a research topic. Furthermore, a conceptual literature review discusses how this understanding was reached and attempts to determine whether a greater understanding can be suggested. It also provides a snapshot of where things are within a particular field of research.

Content of a Good Literature Review

Sources for literature reviews typically include books and journal articles. Because the literature review focuses on the process of locating, reading, and synthesizing materials on a given topic, researchers may first turn to the Tier 1 academic journals in their field, seeking knowledge from the most prestigious sources. However, this should not be the only determining factor in selecting source material. Conference papers may offer valuable information and government reports may also serve as a relevant, credible source. Importantly, without establishing the key points of previous research, it is impossible to establish how the new research advances the topic. The review provides a framework for relating new findings to previous findings. Diligent investigations can also turn up current, relevant, well-researched studies from dissertations, smaller universities, and lesser-known journals. This focus also ensures that the most relevant studies will be discussed in the review and that other less relevant research may be left out. Hart (1998, p. 27) contributes additional suggestions for reviewing the literature, including:

- distinguishing what has been done from what needs to be done
- discovering important variables relevant to the topic
- synthesizing and gaining a new perspective
- identifying relationships among ideas and practices
- establishing the context of the topic or problem
- rationalizing the significance of the problem
- enhancing and acquiring the subject vocabulary
- understanding the structure of the subject
- relating ideas and theory to applications

- identifying the main methodologies and research techniques that have been used
- placing the research in a historical context to show familiarity with state-of-the-art developments.

The literature review "forgoes the chain of reasoning links to past research and helps fashion the problem's rational that contributes to building a credible explanation" (Krathwohl, 1998, p. 101). That is, a literature review examines books, scholarly articles, dissertations, conference proceedings, media, and other resources which are relevant to a particular area of research and provides context by identifying past research and identifying gaps in the current research, justifying the project.

If researchers are prepared to begin the literature search, then it is assumed that they have a well-formulated research question(s) or objective. A clearly articulated objective is the fundamental aspect that informs the type of information needed and the identification of relevant literature. "The research questions provide the structure for the whole of the review" (Jesson et al., 2011, p. 18). Researchers use the research question to help guide the search process and the writing process itself, so the question should be focused, concise, yet complex.

With clarification of the research objective/question(s), attention must be turned to finding credible and relevant research sources. The breadth and depth of source material is colossal. With the turn of the twenty-first century two decades behind us, technology certainly continues to widen our options for a literature search. There are literally hundreds of databases and sources available online. How do we begin? How do we know what is important enough to cite as a resource for a project? Let's put the information we've talked about so far in context. The next section outlines the five key strategies to aid in a successful review of literature.

Five Key Steps: Strategy for a Strong Start

In no way is this chapter a comprehensive presentation of all the different types of literature searches. Primarily, it attempts to explain the five key steps to completing a comprehensive literature search in preparation of performing research. Those steps are: *identification of key terms; relevancy*; *organization*; *synthesis;* and *drafting.*

Because a literature review must provide a research context for a particular topic, the primary objectives are to summarize the body of existing research on a specific topic, to analyze the conceptual content of the field, to identify patterns and themes, and to become informed on the strengths and weaknesses of selected literature. There are thousands of sources to be mined in a literature search. The enormous amount of textual information, mostly unstructured and without any semantic description, necessitates the development of tools that help researchers to effectively and efficiently process huge amounts of data and quickly search, classify, and assess their relevance. Browsing such huge quantities of data is easier if a subset of words describes the main content of the sought-after documents used. The key-term search refers to the querying of scholarly databases with the specific use of words or phrases when attempting to locate relevant literature.

Identification of Key Terms

The initial approach in planning the search strategy is to identify key terms. These terms serve as a highly concise summary of a document and aid in identification and retrieval based on their content. Key terms are used in academic articles to give an idea to the reader about the content of the article (Siddiqi and Sharan, 2015, p. 18). “The identification of the keywords of a research study is the first essential step in identifying relevant literature. Unless this is done in a careful, logical way, you will probably fail to identify some of the key areas of the literature” (Oliver, 2012, p. 129). Key terms, also commonly called *keywords* or *search terms*, represent the main concepts of a research topic and are the words used in everyday vernacular to describe the topic. A *keyphrase* connotes a multi-word lexeme (e.g., healthcare proxy), whereas a keyword is a single word term (e.g., education). Using single words or open compounds as index terms can result in confusion. See the next bulleted section for a detailed explanation of this example. Without the right key terms, it is difficult to find the books and articles needed. It is important to consider the key terms related to a search topic and consequently establish an appropriate vocabulary. However, this could prove difficult if the topic is new to the researcher. “Keyword searching presents a classic cold-start problem for the novice researcher. How can one identify the applicable keywords for an unknown domain? The best source for keywords is, of course, the literature source for the domain. All articles reviewed should be read with an eye for potential keywords” (Levy and Ellis, 2006, p. 190). This is a conundrum. The aim of the key-term search is to identify terms, words, or phrases that describe the subjects of documents in the best possible way. These words can be found by scanning for the research objective/questions first. *It is often helpful to start with key terms and look at a few items from a results list. If relevant hits are found, looking at the list of subject terms in those records is an effective strategy.* Using these terms and running the search again also works well.

Once a few general results are researched, examining them for search terms aids progress. Gleaning through a few foundational articles on the topic will quickly orientate the researcher to keywords. Furthermore, key terms represent the main theme of a text so they can be used as a measure of similarity for text clustering (Siddiqi and Sharan, 2015, p. 18). Always noting the key terms identified in articles when beginning a collection is essential. These strategies will help narrow or expand the search along with keeping a list of key terms in the ongoing academic search.

Databases and Search Engines

Most university libraries have access to hundreds of databases storing scholarly research. Databases can be multidisciplinary, or they can specialize in specific subject areas. There are psychology databases, education databases, nursing databases, etc. Common databases for general searches include ProQuest and EBSCO. JSTOR, APA PsycInfo, and PubMed, to name a few, provide access to more than 12 million academic journal articles, books, and primary sources in 75 disciplines. More than one database for a comprehensive search on a topic must be examined.

Although there may be some overlap, each database typically contains different journals and may provide different results. Furthermore, free search engines such as Google Scholar, World Wide Science, Microsoft Academic, and Refseek may offer results not found elsewhere. Be advised though, any source, regardless of its location, should be thoroughly examined to ensure its academic legitimacy. Once the databases have been identified, the search can begin. *A general rule is to start with broad searches. Cast a wide net and explore the results.* Each search should be repeated on different databases. These tips will be useful at the beginning of a search:

- Breadth. Prioritize the key terms and begin with the two most important concepts. Do not use any limiters initially (e.g., date restrictions and peer-reviewed).
- Boolean operators. Three operators are *AND*, *OR*, and *NOT.* These are combined with key terms into powerful searches:
  - Use AND between words which represent the main ideas in the question: *adolescent AND technology.* This will find results with both search words.
  - Use OR between words that mean the same thing: *adolescent OR teenager.* This search will find results with either (or both) of the search words.
  - Use NOT to exclude words that you do not want in your search results: *(adolescent OR teenager) NOT "young adult."* This search will find results with the desired keywords but with the undesired ones.
- Keyphrase. Search using quotation marks. Double quotation marks help you search for common phrases and make your results more relevant. *"Blood test"* will find results with the words "blood test" as a compound noun rather than "blood" as a noun and "test" as a verb. Many databases automatically insert the Boolean AND when you enter multiple terms. To search for a keyphrase, put the words in quotation marks.
- Truncation. The asterisk symbol * is used for truncation which will help search for different word endings. *teen** will find results with the words: teen, teens, teenager, teenagers.
- Range. Search terms within specific ranges of each other. Proximity searching allows the researcher to specify where search terms will appear in relation to each other. For example, *teenager w/10 digital literacy* will search for *teenager* within ten words of *digital literacy.*
- Suggestions. Some databases will suggest keywords as new keywords are entered into the search box. Make note of the suggestions.
- Revision. Narrow and refine the search results by year of publication or date range (for recent or historical research), document or source type (e.g., article, review, or book), and subject or keyword (for relevance). Try repeating the search using the

'subject' headings or 'keywords' field to focus the search in particular fields such as the citation and abstract.

Though a broad search is recommended as a first step, Krathwohl (1998) reminds us that a wide-ranging search may result in too many results to screen, yet a narrowly precise search may exclude relevant items. A search should cast a net just beyond the immediate boundaries of the target area. Using preliminary searches to narrow in on key terms and combining those search terms make an ideal strategy for the most effective yet inclusive search.

Revising the search and updating key terms is the way to identify the most relevant studies. An exhaustive literature search is the ideal procedure to create a high-quality review. The search should capture as much literature as possible, while inclusion and exclusion criteria will be used to reduce irrelevancies. Nevertheless, it would be impossible to read absolutely everything that has been written on the topic. Yet, before organizing and reading the collected articles, there is a need to ensure an adequate range of relevant literature; doing so ensures that control over the boundaries of the search has been maintained. The initial search of literature could take weeks. Rarely does a researcher jump into the process and narrow down key articles over several days. This process is one of slow discovery and should be allotted the time to explore. A literature search can be a daunting, tiring, and time-consuming task, but because this activity forms the foundation for future research, it is essential for it to be comprehensive.

Relevancy

As the literature search begins to result in a pile of collected studies, an evaluation of which sources are most relevant to the project must be made. The length and number of sources needed for a comprehensive review varies. There is no answer to the question of how many sources should be in a literature review. It is a rare problem that one cannot find enough research to review. Nearly always, the decision is what to cull. Researchers will have to evaluate which sources are most relevant to their project. Undoubtedly, they will be lured by studies that explore avenues of different directions, which may be both distracting and intriguing. B. F. Skinner's enthusiasm should be weighed carefully, "When you run into something interesting, drop everything else and study it" (Skinner, 1956, p. 223). A literature search can quickly go awry if researchers are not prepared to stay focused on the topic and pare down the search results to a manageable, relevant collection. Be open to credible research that may inform the project, but the siren songs of Skinner – shelve those for the next rainy day and future endeavors.

As part of the search, identification of landmarks or classic studies and theorists provide a strong framework/context for study. Identification of seminal research and researchers is the easy part. Wading through a plethora of other research can become complicated. Rather than reading an entire manuscript, a cursory review for relevance and credibility is in order. For each publication, quickly assess the source by asking these questions:

- Where was the research published?
- Has it been peer-reviewed?
- What question is the author addressing?
- What are the key theories?
- What are the methods?
- What are the conclusions of the study?
- How does the research support, add to, or challenge established knowledge?
- What are the strengths and weaknesses of the research?

As researchers begin to read through the gathered material, they should search for common themes as these themes may provide the structure and direction for the literature review. Keep in mind that research is an iterative process – it is not unusual to go back and search information sources for more material.

Finally, most likely, the majority of the research will be accessed online. Researchers must be their own advocate and seek out a research librarian who can save time and help by assisting in the revision of key terms to find relevant information more efficiently. Librarians will give suggestions and give directions to the most suitable databases for a field of study. They can advise on search strategies and techniques tailored to the specific topic and provide referrals to other sources and collections. As the key-term search becomes fruitful, researchers will begin to, rather quickly, collect a growing body of materials.

There is no exact answer to the question of how many articles is enough. Many literature review searches begin with a dearth of research that will suddenly become an abundance. Once researchers begin finding the same studies, with the same authors, with similar citations, search after search, they can feel confident that they are nearing exhaustion of the topic searched. Organizing the research collected is the next step to the literature search.

Organization

A strong literature review is a synthesis of prior research and places the study within a larger body of work. It shows how the study seeks to fill in a gap in or extend knowledge in a topic area. During the literature search, the vast amount of information that is available to researchers is often the cause for anxiety, especially to those who are new to the task. It is tempting to include every relevant study unearthed. That is not the path to scholarship. The challenge is to find the right balance between demonstrating confidence and establishing familiarity with the literature while focusing on the most relevant data for the study at hand.

A practical approach to the organization of the literature search is strongly recommended. There are three stages of organization: initial summary organization; taxonomy of organization; and mapping. These three stages ensure that the literature is organized around a coherent structure for clarity with a logical flow of ideas, organization, and readability.

Initial Summary Organization

Once identified and located, the articles for the review need to be organized in a way that is relevant. Of course, it is impossible to organize the research until a cursory review of each source is made. A classification system is useful as a starting point and for conducting a first summary of research, but an established taxonomy is necessary for the organization and analysis of the primary sources to be included in the review. As each article is found, the steps below outline the initial approach that would be helpful when beginning to read abstracts and summaries of research sources:

- Skim. Researchers need a strong sense of the content of the article without committing a lot of time at this point. Skim the article by focusing on the abstract, introduction, and the conclusion. If is it relevant to the topic, select it.
- Note. Researchers should make clear notes of relevant and key points on the document the first time through. Researchers may revisit a particular document many times, so detailed notes will be valuable in saving time and avoiding having to reread an entire study.
- Log. Researchers should create the complete and accurate citation immediately and identify a topic/theme.
- Manage. Researchers need to manage and organize sources. Reference-management resources use specific tools to help organize the references found during a literature review search. Popular management systems include Endnote, Zotero, and Mendeley. Management systems like these help researchers reassess each piece rather quickly to decide whether to the cut or keep for a close and deep reading as the source number grows.

During the collection process, an initial examination of the literature under consideration may begin to reveal major themes. Bruce (1994) suggests identifying categories and subcategories as early as possible, knowing they can be revised. A fundamental rule of the literature search is to be a good custodian of your research. No one wants to be sifting through piles of papers looking for the one or two interesting articles read three weeks ago. A simple initial spreadsheet created and kept current with each source will help identify the exact content of the growing collection and will provide a reference for easier organization later. See Table 4.1 for an example of the fundamental criteria that should be logged as each relevant research article is collected. It is not prudent to read each entire article as collected. Instead, as suggested above, read the abstract, skim for key aspects, and note the relevant details. As the spreadsheet grows, a pattern will emerge, and a decision can

Table 4.1 *Sample spreadsheet for initial collection*

Citation	Subject/topic	Methods	Strengths/limitations	Key terms	Conclusions

be made determining which articles would be worth advancing to a thorough reading and analysis.

Novice researchers often ask: How many sources do I need? How do I know when I have collected enough? These researchers will read many more research sources than they will use in their final review of literature. That is the nature of the process. Planning to sift through hundreds – even thousands – of pages of documents is impossible without synthesizing the vast amount of research gleaned. Conceptual saturation will be reached when the same themes are repeated and the same citations surface on a regular basis. This is conceptual saturation. It is at this point where it is time to pare down the initial collection into an assortment that is readable within a manageable timeline. Researchers will find that transitioning from the relevant articles in the spreadsheet to a taxonomy will help by organizing the key features of each article.

Taxonomy of Organization

Moving from searching literature to reviewing literature is an exciting transition. But, as explained, not everything collected on the initial search will need to be read. Most likely, at this stage, many literature researchers have been led down various paths and have numerous sources with many subtopics listed in their spreadsheet. Key factors determining whether the collected source makes the “cut” to be read closely need to be identified. Cooper (1988) has developed a taxonomy that classifies literature reviews by six characteristics: (1) focus of attention; (2) goal of the synthesis; (3) perspective on the literature; (4) coverage of the literature; (5) organization of the perspective; and (6) intended audience (see Table 4.2). This chart outlines the taxonomy as a tool to help readers assess the quality of reviews and provides a guiding framework for those conducting their own study.

Hochrein & Glock (2012) recommend the use of Cooper’s taxonomy as a useful strategy for reviewing a singular piece of literature. They synthesize the above categories by expanding on Cooper’s topics. They explain that the *focus* of a literature review summarizes the sub-category findings, research methods, theories, and practices or applications and may not be mutually exclusive. The *goal* of the review is the merging of research findings by resolving conflicts among inconsistent views or by creating a framework to overcome gaps and may be the critique of prior work or the identification of central issues in a certain area. The *perspective* of a review is the manner in which the literature is presented. For example, a review may be neutral or espousal, representing a balanced approach for a certain view. The number of research outlets that were considered in searching for relevant literature refers to the *coverage* of sources. The extent of the coverage influences the number of articles that are included. Coverage may be differentiated into different types including exhaustive (based on almost the entire literature on a topic), exhaustive with selective citation (based on an analysis of a selected sample of works), representative (based on a sample that typifies the larger groups), and central or pivotal (based on efforts that have provided direction for a field). Finally, Hochrein & Glock (2012) explain *organization* as the way a review is arranged

Table 4.2 *Cooper's taxonomy of literature reviews (Cooper, 1988)*

Characteristic	Categories
Focus	Research outcomes Research methods Theories Practices or applications
Goal	Integration (a) Generalization (b) Conflict resolution (c) Linguistic bridge-building Criticism Identification of central issues
Perspective	Neutral representation Espousal of position
Coverage	Exhaustive Exhaustive with selective citation Representative Central or pivotal
Organization	Historical Conceptual Methodological
Audience	Specialized scholars General scholars Practitioners or policy makers General public

and how the content is analyzed; and consider the value of recognition of the intended *audience* and its particular interest in the project. (p. 220).

Thinking about organizational structure early can guide the process as it unfolds, helping in the organization of the sources prior to any analysis. Nearly all literature reviews are structured around major themes or concepts that emerge as the literature is examined and reviewed. Cooper's taxonomy serves well to remove sources that are not credible, reliable, relevant, and, quite frankly, offer nothing new. This culling process will help funnel source material to a small, stronger key focus. As research is organized, read, annotated, summarized, and key factors are noted, a conceptual picture will emerge.

Mapping

Mapping, a classic visual strategy, is a tactic commonly used for organizing source material in literature reviews. Once relevant source material is collected, culled, detailed, and summarized, it is important to get a larger picture of the literature. It is an exciting time to begin to understand the relationships among the research articles. A concept map is a graphic representation of the key ideas of the research and those relationships. Concept mapping is an effective tool that can help make sense of information while conducting a literature review. Concept maps allow users to group information in related modules so that the connections between and among the modules become more readily apparent than they might be from an examination of a list. Rowley and Slack (2004, p. 8) propose "… concept mapping can be a useful way of identifying key concepts in a collection of documents or a research area." They suggest that concept maps can be used as a tool to "… identify additional search terms during the literature search, clarify thinking about the structure of the literature review in preparation for writing the review and understand theory, concepts and the relationships between them" (Rowley and Slack, 2004, p. 8). Moving from traditional written content to mapping may help reveal patterns and connections that are difficult to identify in narrative format. The shift to another modality also helps distinguish links and connections that may otherwise be hidden, identifying gaps in the field. This can be fundamental in helping to identify the parameters of a topic.

To create a concept map, pick out the main concepts of the topic and brainstorm, drawing shapes around the concepts and clustering the shapes in a way that is meaningful to you. Maps may take several different forms including a hierarchical, circular design and flow chart. Heinrich (2001) advises researchers to map as a process of deduction, mapping specific to general concepts (upright triangle shape); or a process of induction, mapping general to specific concepts (inverted triangle shape). Regardless of the design, Creswell (2015, pp. 96–97) provides key advice:

- Identify key terms of your topic and place them at the top of your map.
- Take the information for your map and sort into groups of related topical areas or "families of studies."
- Provide a label for each box.
- Develop the literature map on as many levels as possible.

A map serves as a conceptual model of the collected literature and will help reveal associations to deepen the analysis and synthesis of the research. Hart (1998) stresses that mapping is both an organizational tool and a reflexive tool. This requires metacognitive awareness of understanding the mapping process itself. Maps remove the linear nature of the research and instead positions information in a way that is more natural for the brain to process.

Synthesis

Once the research studies are collected and organized for inclusion in the literature review, the researcher must synthesize the information. Synthesis of research is key to a comprehensive literature review that is well grounded and appropriately places the topic in context. Synthesis demonstrates a critical analysis of the research collected as well as the skill to integrate the results of the analysis into your own literature review. Each article reviewed should be evaluated and weighed for adequacy, appropriateness, and thoroughness before inclusion in the review (Garrard, 2017).

Synthesis entails a deep and though examination of the material for the purpose of integrating, modifying, and generalizing the content in relation to the other sources selected. Torraco (2005, p. 362) highlights the purpose of a comprehensive synthesis:

> Synthesizing new knowledge on the topic with the strengths and deficiencies of a body of literature exposed, authors can take advantage of the breadth and depth of their insights to create a better understanding of the topic through synthesis. Synthesis integrates existing ideas with new ideas to create a new formulation of the topic or issue. Synthesizing the literature means that the review weaves the streams of research together to focus on core issues rather than merely reporting previous literature. Synthesis is not a data dump. It is a creative activity that produces a new model, conceptual framework, or other unique conception informed by the author's intimate knowledge of the topic. The result of a comprehensive synthesis of literature is that new knowledge or perspective is created despite the fact that the review summarizes previous research.

As mentioned earlier, there are services and websites to help researchers with both organization and analysis. A popular resource to use during synthesis is called Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). PRISMA primarily focuses on the reporting of reviews evaluating the effects of interventions but can also be used as a basis for reporting systematic reviews with objectives other than evaluating interventions.

Analysis and synthesis are not a definitive process in the review of literature. Though we highlight synthesis as step four in the literature search procedure, it is rarely a formal process at a set time. "During research, analysis and synthesis are ongoing, interactive, habituated inquiry processes" (Stake, 2010, p. 137). Synthesis and analysis will emerge as a continuous process from the beginning to the final edit of a project. When transitioning from summarizing the content of a source to synthesizing the content, looking for specific connections between the sources and how those relate to the research question is appropriate. Readers need to understand how and why the information from the numerous sources overlap. Synthesis is a way to make those connections among and between numerous and varied source materials.

Moreover, it is important not to confuse summary with synthesis. The key to synthesizing is to extract the most important and relevant information and relate it to similar or dissimilar key points from other sources collected during the literature search, with the goal of providing an overall picture of the state of knowledge on your topic. To synthesize is to combine independent elements and form a cohesive whole picture. In essence, the literature review should integrate your sources and identify patterns among the collected articles – this begins with synthesis.

Drafting

Once the broad decisions have been made about how to organize the literature review, and the analysis and synthesis process are underway, the drafting stage can begin. Consider the metaphor of describing trees verses describing a forest. In the case of a literature review, "you are really creating a new forest, which you will build by using the trees you found in the literature you read" (Galvan, 2006, p. 72). The literature review will rely heavily on the sources read since these sources dictate the structure and direction of the review. It is important that the concepts are presented in an order that makes sense for the context of the research project.

Therefore, another key to a strong literature review is to place each article in the context of its contribution to the topic by describing the relationship of each article to the others under consideration and noting contradictory studies. Identifying areas of prior scholarship prevents duplication of effort and points the way forward for further research. Drawing from the spreadsheet, notes, templates, and maps helps outline the review. Galvan (2006, pp. 71–79) has a useful approach to drafting a literature review as summarized in the bulleted list below:

- Create a topic outline that traces the argument.
    - Explain the line or argument (or thesis); then the narrative that follows should explain and justify the line of justification/argument.
- Note differences among studies within each heading.
- Detail obvious gaps or areas needing more research.
- Flesh out the outline with details from the analysis.
- Describe relevant theories and discuss how studies relate to and advance the theory.
- Summarize periodically and, again, near the end of the review.
- Present conclusions and implications.
- Suggest specific directions for future research near the end of the review.

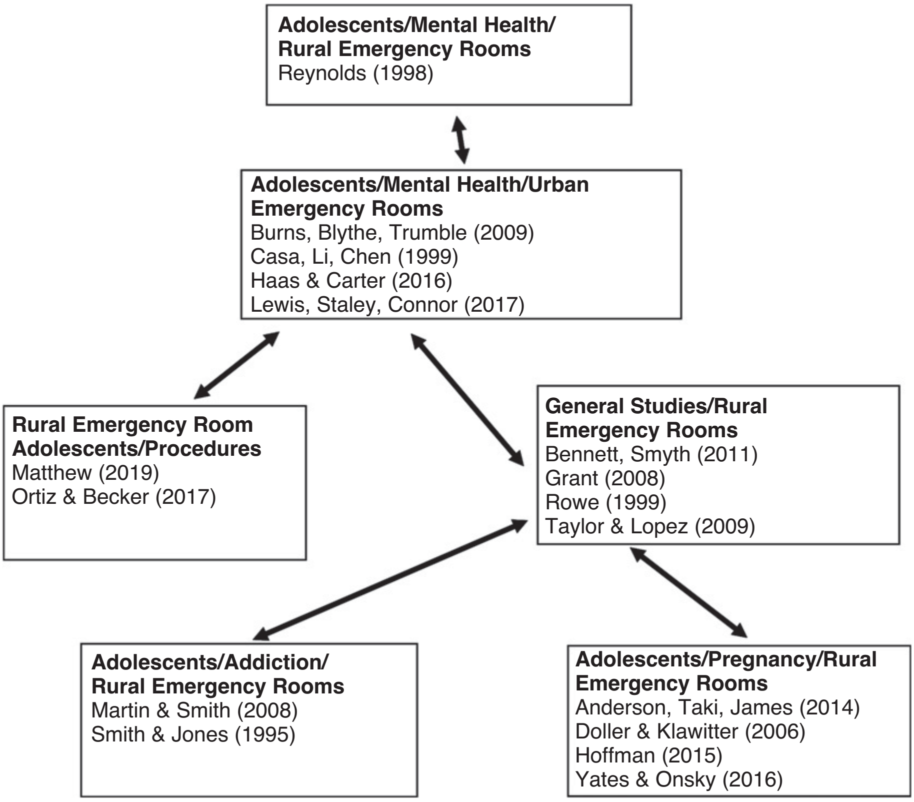
The structure of a literature review, regardless of length, still follows the rules of an essay: introduction, body, and conclusion. The introduction should introduce the topic and give a scope of the review and the organization of the narrative. Each body (i.e., paragraph) should focus on a theme that is relevant to the topic. Synthesis is crucial because more than one source at a time will be discussed. Several of the reviewed readings may be connected under one theme. The conclusion should give a summary of the main agreements and disagreements, gaps, and your overall perspective on the topic. Finally, as with any writing, expect to draft and revise numerous times. The writing of a review of literature is a process. Expect to visit and revisit the manuscript until it is tightened in a logical, coherent review of relevant studies to the topic.

Contextualizing the Five Key Strategies

As this chapter elucidates the different aspects of a literature review, an example will be useful to serve as a foundation for the explanations shared. Imagine Maya, a nurse at a rural hospital in the emergency department that has seen an increase in teens walking in for mental health purposes. Maya has completed her Master of Science in Nursing but sees an opportunity for a research study to help understand and improve practices for teens who are seeking help for mental health issues. There is a grant available for research involving adolescents. She decides to design a study and apply for the grant. Maya's study is called *Rural Emergency Room Mental Health Procedures for Adolescents* (RERMHPA). Maya's literature review will show the grant reviewers that she has an in-depth grasp of her subject, that she understands where her own research project fits into the field, and how that project would fill a gap in an existing body of agreed knowledge.

Maya decides that a systematic approach to the review of literature seems logical. In preparation for research on rural emergency room procedures for adolescents with mental health concerns, Maya starts her search with the key terms: "rural emergency rooms"; "adolescents"; "teen*"; and "mental health." These key terms and phrases will certainly garner a breadth of results that Maya can begin to collect and organize. She realizes that a literature review map would help her organize the collection of research and creates the one shown in Figure 4.1.

At this point, Maya is ready to conduct a close read of her collected research. As her knowledge expands, she begins an outline that traces the argument, noting differences among studies and detailing obvious gaps and areas needing more research. As she drafts the first copy, Maya incorporates the suggestions detailed



Adolescents/Mental Health/ Rural Emergency Rooms
Reynolds (1998)

Adolescents/Mental Health/Urban Emergency Rooms
Burns, Blythe, Trumble (2009)
Casa, Li, Chen (1999)
Haas & Carter (2016)
Lewis, Staley, Connor (2017)

Rural Emergency Room Adolescents/Procedures
Matthew (2019)
Ortiz & Becker (2017)

General Studies/Rural Emergency Rooms
Bennett, Smyth (2011)
Grant (2008)
Rowe (1999)
Taylor & Lopez (2009)

Adolescents/Addiction/ Rural Emergency Rooms
Martin & Smith (2008)
Smith & Jones (1995)

Adolescents/Pregnancy/Rural Emergency Rooms
Anderson, Taki, James (2014)
Doller & Klawitter (2006)
Hoffman (2015)
Yates & Onsky (2016)

Figure 4.1 *Maya's literature review map.*

above. Her first draft describes relevant theories, synthesizes the collected research, presents implications, and discusses how the studies relate to her project. Maya will revise and fine-tune her draft until she has a final copy that is cogent, succinct, yet comprehensively supports her proposed study, justifying the application for her research grant.

Conclusion

The review of literature is a means for researchers to “join the conversation” by providing context, relevant knowledge, concrete methodology, a fresh analysis, and conclusions based on a thoughtful synthesis of their research within the field. A literature review is not just about reporting facts; it requires careful consideration of researched studies to construct an unbiased narrative supported by published evidence. Upon completion of the literature review, a researcher should have a solid foundation of knowledge in the area and a good feel for the direction any new research should take. In fact, the process of reviewing literature often reveals gaps in the field, inconsistencies in findings, and opportunities for deeper exploration. Researchers can build on their own discoveries and file away areas for future exploration. Maya, our emergency room nurse, not only would have collected a strong body of research to justify her application for a grant but, through this process, also she would certainly have been better informed and most likely noted gaps in the field ripe for further investigation.

Considering the increased use of evidence-based practice and research generating stronger evidence (Lyden et al., 2013), literature reviews have become crucial tools for critically appraising prior knowledge in all fields. The review of literature provides a synthesis of research of the given topic to establish context on the subject and to establish researchers’ own position regarding the existing field of scholarship. It is an essential tool for integrating and critically appraising prior research and reveals where the reviewer is entering the academic conversation on the specific topic.

In this chapter, I hope to ease the anxiety for those tackling a review of literature. Specifically, the chapter provides an assessable guide to writing a comprehensive review of literature that begins with a discussion of the purpose of a review and then focuses on the five major components of a successful review: key terms; relevancy; organization; synthesis; and drafting. The review of literature is an opportunity for scholars to demonstrate that their own research draws from and grows out of existing theories and dependable research. It establishes credibility on the topic. Furthermore, a literature review offers a fresh perspective that leads to a researcher’s own contribution to a growing body of knowledge. When rigorously and systematically conducted, the literature review is an opportunity to share critical perspectives, to establish the importance of a topic within the broader academic community, and to present important information for scholars and practitioners looking for state-of-the-art evidence in each area of inquiry.

References

Boote, D. N. & Beile, P. (2005). Scholars before researchers: On the centrality of the dissertation literature review in research preparation. *Educational Researcher*, *34*(6), 3–15.

Bruce, C. S. (1994). Research students' early experiences of the dissertation literature review. *Studies in Higher Education*, *19*(2), 217–229.

Cooper, H. M. (1984). *The Integrative Research Review: A Systematic Approach*. Sage.

Cooper, H. M. (1988). Organizing knowledge synthesis: A taxonomy of literature reviews. *Knowledge in Society*, *1*, 104–126.

Creswell, J. (2015). *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research*, 5th ed. Pearson Education, Inc.

Galvan, J. (2006). *Writing Literature Reviews: A Guide for Students of the Behavioral Sciences*, 3rd ed. Pyrczak Publishing.

Garrard, J. (2017). *Health Sciences Literature Review Made Easy: The Matrix Method*. Jones and Bartlett Learning.

Hart, C. (1998). *Doing a Literature Review: Releasing the Social Science Research Imagination*. Sage.

Heinrich, K. T. (2001). Mind-mapping: A successful technique for organizing a literature review. *Nurse Author & Editor*, *11*(2), 7–8.

Hochrein, S. & Glock, C. (2012). Systematic literature reviews in purchasing and supply management research: A tertiary study. *International Journal of Integrated Supply Management*, *7*(4), 215–245.

Jesson J., Matheson L., & Lacey, F. M. (2011). *Doing your Literature Review: Traditional and Systematic Techniques*. Sage.

Krathwohl, D. R. (1998). *Methods of Educational and Social Science Research: An Integrated Approach*, 2nd ed. Longman.

Levy, Y. & Ellis, T. J. (2006). A systems approach to conduct an effective literature review in support of information systems research. *Informing Science: International Journal of an Emerging Transdiscipline*, *9*, 181–212.

Lyden J. R., Zickmund S. L., Bhargava T. D., et al. (2013). Implementing health information technology in a patient-centered manner: Patient experiences with an online evidence-based lifestyle intervention. *Journal for Healthcare Quality*, *35*(5), 47–57.

McCabe, T. (2005) How to conduct an effective literature search. *Nursing Standard*, *20*(11), 41–47.

Mertler, C. A. & Charles, C. M. (2011). *Introduction to Educational Research*. Pearson/Allyn & Bacon.

Moher, D., Shamseer, L., Clarke, M., et al. (2015). Preferred reporting items for systematic review and meta-analysis protocols. *System Review*, *4*, article1.

Nobre, M. R, Bernardo, W. M., & Jatene, F. B. (2003). Evidence based clinical practice. Part 1 – well-structured clinical questions. *Revista da Associacao Medica Brasileira*, *49*(4), 445–9.

Oliver, P. (2012). *Succeeding with Your Literature Review: A Handbook for Students*. Hill Education.

Rhoades, E. A. (2011). Literature reviews. *The Volta Review*, *111*(1), 61–71.

Rowley, J. & Slack, F. (2004). Conducting a literature review. *Management Research News*, *27*(4), 31–39.

Siddiqi, S. & Sharan, A., (2015). Keyword and keyphrase extraction techniques: A literature review. *International Journal of Computer Applications*, *109*(2), 18–23.

Skinner, B. F. (1956). A case history in scientific method. *American Psychologist*, *11*, 221–233.

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, *104*, 333–339.

Stake, R. E. (2010). *Qualitative Research: Studying How Things Work*. Guilford Press.

Torraco, J. (2005). Writing integrative literature reviews: Guidelines and examples, *Human Resource Development Review*, *4*, 356–367.

Webster, J. & Watson, R. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, *26*(2), xiii–xxiii.

5 Choosing a Research Design

Glynis M. Breakwell

Abstract
A research design is the sequence of things done in order to collect the information needed to answer a research question. The design states which data will be sought, from which sources, at which times, and in which ways. This chapter describes the influences that shape the decisions researchers must make when constructing a research design. Research designs differ in the source of information used, whether data used are naturally occurring or a result of intervention, and the way data are elicited, recorded, and analyzed. Typically, the nature of the research question, assumptions on which it is based, and ethical considerations drive the design construction. I describe seven major influences on design choice in this chapter: research question novelty; levels of analysis and explanation used; epistemological and ontological assumptions; characteristics of data sources; data analyzability; piloting results; and various practicalities. Understanding these influences will improve research design decisions.

Keywords: Levels of Analysis; Data Collection and Analyzability; Piloting; Ontology; Epistemology; Design Practicalities

Introduction

Deciding upon a research design is rarely a matter of selecting something that is prefabricated. Instead, it entails constructing a customized package of activities that are aimed at answering a specific research question. Choosing a research design is a challenging and creative enterprise, swayed by many influences. Every researcher should be aware of these influences. This chapter is structured to systematically present the main choices that have to be made and this brief introduction summarizes the main issues that will be covered. In particular, the chapter outlines how types of research design are differentiated along the following dimensions:

- the source of information used
- whether data used are naturally occurring or a result of intervention
- the way data are elicited, recorded, and analyzed.

However, in practice, such types of design have been amalgamated and hybridized. The nature of the research question, and the theoretical or interpretive assumptions on which it is based, drive the construction of the design. The design will also be constrained by ethical considerations (see Chapter 2 in this volume). In all, I present seven important influences on designing a research study:

- When the research question is fundamentally new (perhaps because of changes in the physical or social context) and the research is exploratory, the design cannot be based uncritically on extant templates. The absence of useful precedents encourages design innovations to emerge.
- Designs are influenced by the levels of analysis and explanation that interest the researcher. Explanatory and predictive models can be lodged at different levels of analysis – the neurological, intra-psychic, inter-individual, group and intergroup, societal, and inter-societal (including ideological and social representational systems). The level chosen will influence which sources and forms of data are used in the design.
- Any epistemological and ontological predilections the researcher has shape design decisions. Translation of these philosophical assumptions into positivist, neo-positivist, constructionist/interpretational, or critical–ideological methodological approaches guides design construction.
- Once the researcher knows which types of data are needed, the design depends on decisions that have to be made about the characteristics of data sources, the techniques for accessing the data, and the form in which data are recorded.
- Design construction will need to ensure that the data to be collected can be analyzed appropriately and in the manner preferred by the researcher. Typically, this will affect sample sizes and levels of measurement used.
- Designs should be piloted (i.e., all elements tested on smaller scale, to establish whether they work in the way anticipated, prior to the final roll out of the research). Piloting often reveals weaknesses in the proposed design. Revision of the design is then required. Rigorous piloting can significantly improve the design.
- Practicalities constrain a researcher's choices when constructing the design. These practicalities include time and timing issues; financial limitations; research funding agency priorities; the skills available in the research team; the pressure of scientific opinion brought to bear through prevailing methodological orthodoxies; the preferences of journal publishers, reviewers and editors, and the institutional hierarchies that determine professional advancement and recognition; and the researcher's own habits and reputation.

By the end of this chapter, the reader should be alert to the choices that have to be made in constructing a research design. It is all about knowing. In your research:

- know what you want to know
- know what is already known
- know which data elicitation and analysis methods are available to you
- know how to use these methods ethically
- know the practical constraints upon you
- know where the limits of your own knowledge and skills lie.

And, when you do not know, find out.

Aspects of a Research Design

A research design is the sequence of things that you do in order to collect the information (often just called ‘data’) that you need to answer your research question (Breakwell et al., 2020, p. 14). The design is simply the structured plan for the research. It states which data you will try to get, from which sources, at which times, and in which ways. It details how data will be recorded. It also helps if you know, prior to starting their collection, how you intend to analyze the data that you get.

It is valuable, when beginning to construct a research design, to record your decisions. The decision-making is part of the design process. From the start to the completion of a study it is good practice to maintain a contemporaneous record of what decisions you are making about the research question you are addressing, the way data are accessed, and forms of analysis undertaken. One advantage of doing this is that it can minimize unrecognized “ mission drift.” Mission drift is what happens when a researcher loses sight of the initial research question as a study progresses. Sometimes this happens because the original hypotheses start to look less justified or, perhaps, the data that are accumulated suggest new directions for investigation. Whatever the reason, mission drift is not uncommon. In practice, it may not be a bad thing if it results in unanticipated discoveries. However, it is important for researchers to recognize and acknowledge that the changes in the research design have occurred and to look at their implications for the validity of the theoretical models that they are proposing. Recording all details of the research design and the analyses undertaken allows mission drift to be assessed. There is a further reason for undertaking this contemporaneous recording. Open-access and transparency guidelines in the social and behavioral sciences now often require or strongly recommend that data are made available to other researchers and in a form that makes them open to reanalysis. Clear recording of all details of the construction and execution of the research design and the analyses undertaken is consequently increasingly a necessity.

The idea of a research design runs counter to, but perhaps parallel with, the rocket scientist Wernher von Braun’s assertion that ‘Basic research is what I’m doing when I don’t know what I am doing’ (von Braun & White, 1953). The research design is a premeditated specification of how you will go about finding out what you need to know and how you will make sense of what you find. It does not assume you already know what you will find, yet it is based on the assumption that you have an idea of how you should go about finding it. It does not preclude serendipitous discovery nor, indeed, flawed data or failure. In that it allows for chance Eureka moments, having a research design may not be so different from von Braun’s approach to basic research. As Carl Sagan once said ‘somewhere, something incredible is waiting to be known.’ Research designs may just make the waiting shorter.

Types of Designs

Research designs are often labeled in terms of “types.” Several types are described in Section 3 of this handbook. Generally, the types have been differentiated on a number of dimensions. The first is the nature of the sources of information that you use. The source can be primary or secondary (or even further removed). For instance, if you want to know about a person's health, you can ask or observe her (primary), ask people who know her (secondary), or, perhaps if access is permitted, look at her medical history on an official record (held, for instance, by her health care provider or insurer). The point here is that the object of your enquiry is not necessarily the source of your information. This is particularly likely when you are interested in information about groups or social categories rather than individuals. One important consideration in choosing the source for your information is how reliable (i.e., truthful and/or accurate) the source is. It is clearly important to use sources that are very reliable. If possible, the reliability of a source should be checked or tested. Cross-tabulating data from different sources can be used to pinpoint discrepancies that suggest a lack of source reliability. Alternatively, reliability can be assessed by asking for the same data from a source on several occasions or in several different ways. Inconsistencies in the data generated may then indicate unreliability. There are many reasons why a source may be unreliable (or become unreliable over time) but they fall into two broad clusters: seeking to misinform and being ill informed. The increasing importance of online data collection, where sources may be anonymous and untraceable, emphasizes the need to be cautious about source reliability.

The extent to which you deal with naturally occurring data or something that you have intervened to influence is another important consideration. For example, if you want to know how people's behavior changes when they are under stress, you might collect data in situations that are spontaneously stressful, or you might artificially induce stress in a sample of respondents by manipulating their experiences. Different types of research designs involve different levels of control of the respondent's experiences. Field studies in natural contexts sit at one end of the spectrum with fully structured experimental studies at the other. Between these two poles, there are many variants (including quasi-experimental approaches) each involving features that the researcher can control.

Researchers must also decide how data are elicited, recorded, and analyzed. The great delight of research is that the same research question can be approached in many different ways. For instance, you might ask a source to give information about themselves in writing, verbally, audio-visually, online, or face to face. The choice of the medium may shape the options for recording the information and its subsequent analysis. It also is likely to affect the sorts of information that will be available to you. For example, a written response precludes access to the variety of non-verbal cues that might accompany a face-to-face exchange and influence your interpretation of the verbal message. The idea that the same research question can be addressed using different methods ties into the ongoing debate about the importance of replication of findings. Nosek & Errington (2020) argue that the common assumption that replication entails repeating a study's procedure and observing whether the same results occur is misguided. Arguing that the purpose of replication is to advance theory by confronting existing understanding with new evidence, they suggest that the value of replication may be strongest when existing understanding is weakest. Successful replication provides evidence of generalizability across the conditions that inevitably differ from the original study. Unsuccessful replication indicates that the reliability of the finding may be more constrained than recognized previously. The focus upon generalizability as an alternative to replicability has been important for improving the way some research design types (e.g., observational field studies), which are not capable of complete replication, are accepted and valued.

Traditionally, research designs were differentiated by whether and to what extent they sought to quantify data. Qualitative research and analytical approaches are discussed in Chapters 20 and 28 of this volume. It is worth saying here that choosing the levels of quantification to deploy is a major decision in designing any study. However, it is also worth adding that qualitative and quantitative methods are often now used in concert to address a research question. The design of a study may well include qualitative and quantitative components. The issues this raises for arguments about the philosophical underpinning of method choice are considered briefly later in this chapter.

Taking into account these dimensions, research design types include:

- case studies (involving data from one person or a single discrete community such as a school)
- cross-sectional studies (data collected from a series of separate defined sources at about the same time)
- longitudinal studies (with data from the same person or community over a period of time or over a series of extended time points)
- time series studies (involving collecting data on the same variables at different time points from different sources)
- experiments (involving data collected following researcher manipulation of inputs to a person or a social system and where extraneous influences can be minimized, largely through randomization and sampling strategies)
- quasi-experiments (in these the researcher has limited or no control of the manipulation, over who is manipulated, and/or who is assigned to each manipulation).

However, in practice, such “types” get amalgamated and hybridized. For example, the longitudinal cohort sequential research design is a valuable tool that pulls together cross-sectional and longitudinal elements. This entails collecting data at a series of time points from cross sections of respondents drawn from a number of different age cohorts. The design is used typically to tease out the relative effects of chronological age and changes in socio-economic context or structural factors over time upon cognition and behavior (Breakwell & Fife-Schaw, 1994). Sometimes, longitudinal cohort sequential designs also embed experiments within the data collection. The scope for imaginative mixing of design formats is enormous. Given the malleability and responsiveness of types of research design, increasingly a researcher will not simply choose some pre-existent model without modifying it at least to some degree. It becomes less a matter of choosing a research design, as it were off the shelf, and more one of tailoring, customizing, and constructing it.

The advent of data science and data analytics is particularly challenging the usefulness of thinking about discrete research design types. The “data deluge” (the enormous and often uncontrolled flows of data from innumerable information-sensing devices now available) is open to being made intelligible by artificial intelligence, reliant upon complex statistical models. Research design for the social and behavioral sciences could just become a question of choosing which streams of the deluge to channel toward analysis and structural modeling. However, this would rest upon having confidence in the validity of the initial data and the reliability of the database system compiling it. This is a big act of faith given the scope for both error and cyber corruption. It does suggest that researchers who follow this path will probably gravitate toward multidisciplinary teams with wide-ranging skill sets (probably across the mathematical and statistical sciences, computer science, neuro-cognitive, and social sciences). It seems very likely that the formation of such multidisciplinary teams will promote the formulation of complex research questions that lie at different levels of analysis and explanation. This complexity will precipitate multi-layered research designs that cut across the old typology.

Influences on Choice or Construction of a Research Design

Several factors that should influence the researcher's choice or construction of a research design are examined next. These are in addition to the overarching requirements of conducting research ethically. Since research ethics are fully examined in Chapter 2 of this volume, they are not a focal consideration in this chapter. However, it is important to emphasize here that the design for a study must always comply with the ethical standards and guidelines established for the social and behavioral sciences, and the researcher's choices should be constrained by ethical considerations. In addition, there are other factors that one should consider:

- How exploratory is the research?
- The levels of analysis and explanation inherent in the formulation of the research question.
- The researcher's epistemological and ontological predilections.
- The types of data required: sources, elicitation, and capture.
- Assuring analyzability: levels of measurement determine forms of analysis.
- Does the research design work? Pilot and change.
- The practicalities that constrain choice.

We will examine each of these influences in turn.

How Exploratory Is the Research?

In choosing a research design, it is helpful to consider whether your research question is fundamentally new. Imagine a scenario where you know your research question, have done your literature review, have identified what is known and what are the "known unknowns" about the topic, and may even have glimpsed the shape of the "unknown unknowns" about it. For some questions, there may be little previous research to rely on. This may happen when changes in physical or social contexts crystalize new research questions. For instance, when a global pandemic strikes (as it did with the advent of the coronavirus SARS-CoV-2 in 2019), some of the questions asked by social and behavioral scientists about policy interventions to control the spread of the disease and for managing its aftermath will be new. Comparisons with what had been done in research during other infectious disease outbreaks are made but there is also a need for developing research questions and designs specifically tailored to the context and exigencies of the new pandemic (e.g., the significance of the emergence of more dangerous variants of the virus). In such circumstances, the researcher is exploring the terrain. The research designs used must respond to the actual changes in the situation within the pandemic.

In such a fast-moving and unpredictable situation, accepting that the design itself will be conditional and exploratory is necessary. When the research questions are genuinely new and having to respond to contextual change, it may be useful not to opt for only one design. It may be helpful to use a series of designs, each relevant to a different phase in the unfurling of the crisis. Using several different methods for collecting data simultaneously or sequentially can be a safety net in such exploratory work. Social and behavioral scientists seeking to conduct socially relevant research, based upon contemporaneous evidence, will often find that the research designs they use will be multiplex (cutting across established types of design) and flexible (evolving during the course of a project).

While some research questions are fundamentally new, most are not. Predominantly, they derive from a long history of previous research. In many cases, they derive from well-established theoretical models or propositions. The process of conducting a systematic literature review is designed to reveal how your initial formulation of the research question relates to what has been done before and what has already been reported (see Chapter 4 in this volume). The process allows you to refine the question and understand how it relates to existing theories or interpretive frameworks. It also tends to ensure that you know how previous researchers have designed their studies. As a result, when you come to choose your own design, you have a clear picture of which methods have been typically used. This can help you to avoid those that have proven fruitless and focus upon those that have been productive. However, this has a downside. It can quell innovation. Following a well-tried formula for examining a research question may increase the likelihood that your data will fit an existing body of literature, but it may also limit your ability to offer something new that challenges dominant theoretical models. There is a balance to be struck between following the design norms for research on a topic and introducing alternative designs. Progress in science is dependent upon using what has been discovered in the past but not being constrained by it.

This suggests that it is always vital to evaluate the strengths and weaknesses of the research designs that have been used in the literature that are pertinent to your own question (see Chapter 4 in this volume). This assessment may cover at least two aspects. First, is the design actually capable of addressing the research question posed? For example, is a cross-sectional questionnaire survey capable of revealing how sexual identifications change over the life span? It might tell you something of differences between age cohorts but nothing about changes in one person over time. It is notable that sometimes the limitations of the chosen research design will cause the research question to be retrospectively reframed.

Second, has the design been executed correctly? When choosing a design, the first question is most important. The answers to the second might be significant if the failure to execute is a function of the structure of the design itself rather than merely inadequacy on the part of the researcher. For instance, this might occur if the design is based upon the premise that people will participate in an experiment and, in fact, no one will, or, more likely, the people who are most salient for the research question refuse to participate. Such refusal may be determined by the topic of the research, in which case the design is inappropriate because it is unworkable in practice. The issue of the design that proves unworkable is revisited later in this chapter when discussing the need to pilot any design.

In conclusion, there are two simple warnings for anyone choosing a research design. If your research question is fundamentally new (i.e., not previously researched or previously inadequately researched) be ready to diversify across research designs. If your research question is built upon earlier research, ensure that you learn lessons from the strengths and weaknesses of the designs used before and never, uncritically, adopt someone else's design.

Research Questions and Levels of Analysis and Explanation

The previous section talked a lot about the research question. However, the research question concept is complex. In fact, the chapters in this volume leading up to the current one have discussed the theoretical basis, prior literature, and creative thought process that results in the derivation of a research question. The research question should drive the choice of the initial template for the research design, but it will also shape the modifications that you introduce to any basic template to achieve the answers that you seek. At the forefront of your mind, when you develop a research design, should be the thought (with due acknowledgement to the Spice Girls), "What do I want to know; what do I really, really want to know?"

In most social and behavioral research, the answer to this challenge has to be unpacked carefully because parts of the answer are nested inside a series of other questions (like each doll in a set of Russian dolls). This is because we build explanatory and sometimes predictive models that cut across different levels of analysis – the neurological, intra-psychic, inter-personal, group and intergroup (i.e., social category position), and societal and inter-societal (including ideological and social representational systems) models (see Doise & Valentim, 2015). For instance, the initial research question might be: Why do some people refuse to be vaccinated against a virus that has generated a global pandemic? This question may be answered at many levels of analysis. The possibilities include the intra-psychic – some people believe that they are not at risk from the virus, they fear the possible side effects of the vaccine, or they do not believe the vaccine will be effective. These explanations arise from a level of analysis that is located in the individual's beliefs and feelings. Equally, an interpersonal or group level can be the focus – some individuals are part of networks of conspiracy theorists that promote rejection of the vaccine, or they are members of social categories that generally feel disadvantaged and mistreated by the medical establishment such that they do not trust the motive for vaccinating them. Aligned to the group or social category level of analysis, there are explanations that center upon societal structures and their influence – differences in vaccination acceptance are then deemed a product of differential education, poverty, religion, social media exposure, etc.

It does not take much imagination to realize that the level of analysis and explanation that you, the researcher, prefer could massively influence the research design that you develop. Hence, being clear from the start about the levels of analysis and explanation that you will use is essential. This does not mean that you have to be restricted to only one level. You can self-consciously decide to range across anything from the neurological to the inter-societal. In fact, the answers to important research questions in the social and behavioral sciences rarely sit squarely and solely in one level of analysis. Some theorists argue that a generic framework for the development of theories of social action must include explanations that lie on a variety of levels (Breakwell, 2014). Certainly, trends in the use of "big data" and growing dependence upon artificial intelligence and data analytics would suggest a move in this direction.

The Researcher's Epistemological and Ontological Predilections

Preferences between research designs are driven, to some extent, by the epistemological and ontological assumptions that the researcher makes. Epistemology is the field of philosophical inquiry that investigates the nature, origins, scope, and justifications of knowledge. It delves into how knowledge is acquired. Ontology is the study of the nature of being, existence, and reality. Put very simply, philosophical consideration of the nature of reality has led to two traditions of thought: realism (that argues there is an objective reality to be discovered) and relativism (that posits there is no objective reality, and that instead there are multiple realities that are constructed through interpretation). The four main research traditions in the social and behavioral sciences based on these philosophical debates are positivism, neo-positivism, constructivism–interpretivism, and critical–ideological approaches.

A researcher's adherence to one of these approaches may strongly influence decisions about research design. For instance, positivism traditionally has been associated with quantitative methods and the inductive inference from data of generalizations leading to theory building. In contrast, neo-positivism is associated with the hypothetico-deductive (or scientific) method, according to which the researcher starts with a theory, formulates hypotheses (i.e., predictions) based on this theory, and then collects data to 'test' (i.e., falsify) these hypotheses. Both positivism and neo-positivism accept the existence of an objective reality. They differ in their epistemological premises – neo-positivism challenges the primacy of the senses as the route to acquiring knowledge and instead emphasizes the role of critical rationalism (Blaikie & Priest, 2019). Neo-positivism is particularly associated with the experimental method.Constructivism–interpretivism is founded upon the assertion that there are multiple realities and that these can be known by examining lay meanings and interpretations accessed through detailed, in-depth, and contextualized research. This approach essentially assumes that there is no reality independent of human interpretation. It is socially constructed through the actions of people. As a result, idiographic research designs, such as case studies, are employed while reductionist and quantitative approaches are mostly rejected (Bakker, 2010). Critical–ideological approaches go further. They seek to challenge and transform society. They argue that there are multiple realities mediated by the power relations that are socially and historically constructed. Ponterotto (2005) gives some feminist and Marxist research as examples of this approach. The research designs favored within the critical–ideological approach are similar to those used generally by constructivists.

Ontological and epistemological issues may not be consciously uppermost in your mind when you choose a research design. You may be agnostic about such questions. However, it is worth considering the philosophical predisposition that you rely on to inform the formulation of your own research question and choice of your design. By importing ideas from previous research, you can be embedding its philosophical assumptions in your own.

Once you start thinking about the philosophical assumptions underlying the processes of research design, it is worth considering a word of caution. In practice, clear distinctions between the four philosophical approaches in the forms of research design they inspire tend to disappear. Some heavily experimental research has strong ideological and critical objectives. For instance, early psychological studies examining the role of nature and nurture in the development of gender differences often involved the use of structured experiments, but they were also founded in a critical–ideological approach to nature-based theories of differences between people biologically identified as male or female (see Archer & Lloyd,2002). On the other hand, some qualitative research methods are commonly used by researchers who behave as if they are grounded in a positivist reliance upon inference from their data to build their interpretive frameworks (which some might call theories). For example, researchers employing thematic analysis (Jaspal, 2020) of qualitative material will actively strive to use purely inductive methods to interpret their data (trying as far as possible to exclude their own prior preconceptions even though, as Clarke & Braun (2014) point out, this can be difficult to do).

The prime takeaway message from this short foray into the philosophical underpinnings of research design is that researchers should reflect on their own epistemological and ontological beliefs but should not expect that these will automatically drive their decisions about research design. Many influences besides philosophical predilections will shape the choice of research design.

Types of Data Required: Sources, Elicitation, and Capture

The most basic influence on the choice of research design is clearly which data are needed to address the research question? Assuming that you have refined the question to its quintessential parts and that you know which levels of analysis and explanation you wish to use, you should have a list of the data that you need. The task then falls roughly into three parts that each involve many decisions. The first decision to make is to decide how will you access the data source. You will need to specify the characteristics of the sources (this applies irrespective of the type of sources, whether they are individuals, communities, archives, online repositories or real-time feeds, societal institutions, etc.). You will need to decide how you will select particular sources for the research. The research design should be explicit about the selection rules you use.

Next, you must decide how will you access the data (the elicitation). You will need to decide who will collect the data. It may not be a direct approach from you as the researcher. You may choose to use an intermediary to facilitate the availability or validity of the data. What will be the medium or channel for data collection (e.g., face-to-face, telephone, virtual, textual, verbal, and audio-visual)? To what extent will the source be made aware that data is being collected? In some designs, the source may be ignorant about the data collection (e.g., in observations of behavior of anonymous individuals in public venues). How will you manage gaining permission for data access when agreement is required or preferable? How much will you set out to control or influence the source's willingness or ability to provide data? Persuading a source to co-operate is sometimes a tricky business. For instance, some sections of the public may be reticent or suspicious about giving you data. Equally, some research topics will arouse non-participation or dissembling. Indeed, some data are intrinsically difficult to access. Research designs have to anticipate such difficulties and compensate for them (Breakwell & Barnett, 2020). A corollary issue concerns how much you seek to constrain or manipulate which data the source provides. If you reject or wish to minimize manipulation, some research designs are simply not options for you. You will also need to determine how many data, over what timeline, you want to access. If the period is extensive or a source is involved repeatedly, it is important to attempt to minimize any unwanted effects of long-term or repeated data sweeps. Sometimes this is achieved by designing the sequencing of data collections across sub-samples of sources within a study so that potential artefactual effects are "smoothed" over the entire sample. Such techniques are most frequently applied in experimental or quasi-experimental designs (Fife-Schaw, 2020; Hole, 2020). Whatever your own methodological or theoretical biases are, your decisions in regard to accessing data should always be guided by ethical and legal guidelines for research that have been laid down (see Chapter 2 in this volume).

Finally, you must decide how will you record the data (the capture). Data in the social and behavioral sciences come in many forms. They range from marble cultural artefacts to functional MRI scans, from questionnaire scaled responses to intergenerational oral histories, from neuronal reaction times to stock-market value trends over decades, and many, many more. Whatever their form, data have to be made interpretable. This typically entails using some method for summarizing and systematizing the description of the data. For a single data set, a variety of methods can be used to do this, and in some cases, several methods can be applied with equal justification. The researcher has to decide what method will be used. This is an integral part of the research design.

The decision can be made in advance of data being collected and this may result in the data being structured in a particular way at the point it is collected. For instance, deciding to use self-report on numerical rating scales to collect data on attitudes represents the first step on a path to later quantitative and statistical interpretations of the data collected. Predetermination of the analytic methods to be used in data interpretation is common because of the initial structuring of data during its collection. However, it is not inevitable. Data can be amenable to alternative interpretive approaches because they are not pre-structured by prior design. For instance, data drawn from interviews where verbatim records of interviewee statements are taken may be subjected to many interpretive approaches. One might entail quantification of the number of times a particular phrase is uttered. Another could describe the narrative discourses that emerge across interviews – these can be catalogued and their interrelationships charted using a corpus linguistic approach (Semino, 2017).

Any researcher has two decisions to make. First, how far should they shape the data by the form in which they are initially collected and, thereby, predetermine the methods of interpretation and analysis available. Second, after they are collected, to make them interpretable, how far should they "reduce" the data by creating or deploying a pre-existent conceptual or theoretical framework. Such frameworks are argued to legitimize selectivity in the use of data.

Once you start thinking about these questions, it is evident that the research design is founded upon the battery of decisions that the researcher makes. The decisions will vary in their level of detail or significance. However, they add up to the choice of design because they are the foundation for the construction of the research design that is unique to a specific study. The overall design will then specify the sources of data, the way data are accessed, and how they are captured (and systematically summarized) to make them address the research question.

Assuring Analyzability: Sample Characteristics and Levels of Measurement

The process of constructing a comprehensive research design should always include a consideration of the analytical approaches that will be used once data are collected. This is necessary because each analytical approach has its own requirements of the scale and/or structure of the data. For instance, some statistical procedures will only yield reliable results if the number of sources sampled relative to the number of variables measured exceeds some criterion figure. While statistical power analyses (Cohen, 2013) that establish this criterion can be done retrospectively once data are collected, it is always preferable to calculate the size of the data set required for the desired statistical approach as part of the initial research design process.

Similarly, statistical analyses differ in the assumptions they make about the level of measurement used when capturing data. The traditional typology of levels of measurement distinguishes between nominal, ordinal, interval, and ratio scaling of data (Stevens, 1946). Levels of measurement are discussed in other chapters in this volume. It is sufficient to say here that, while there has been much debate about the usefulness of the typology in relation to the choice of statistic (Gaito, 1980), the level of data scaling will influence which statistics can be legitimately used. The choice of levels of measurement will be the key determiners of subsequent data analyzability. As part of the research design, it is important to be self-conscious in choosing the way data are scaled. In some cases, of course, such as in some qualitative designs, no data scaling is introduced. In such cases, it is equally important to register the intentional absence of measurement in the research design.

Qualitative designs introduce their own assumptions about which data are acceptable in the process of analysis. The various approaches to qualitative data analysis have developed their own interpretive techniques (normally derived, to a greater or lesser extent, from a broader set of epistemological or ontological premises). Such approaches include narrative analysis, discourse analysis, content analysis, thematic analysis, framework analysis, grounded theory, or corpus linguistic analysis). Most involve some form of coding of data to identify common themes, patterns, and relationships. Examination of data sets for the absence of elements that might be expected or the presence of things that are surprising is also common. A research design for a qualitative study could be expected to incorporate a clear statement of how data will be analyzed. Particularly, it could include any a priori theoretical or conceptual framework that will be used as the basis for coding frames. It could be explicit about how the validity of the analysis can be assessed (Yardley, 2000). Articulating, as part of the initial research design, the process through which qualitative data will be interpreted should improve the chances that useful data will be collected. Also, pre-specifying the basis for determining the quality of the analysis may encourage rigor in the choice of sources and methods for eliciting data.

Where the design involves secondary data, collected by someone else (perhaps for a different purpose), control over the form and level of measurement of data is inevitably limited. The extent of these limitations needs to be considered before embarking on the use of any corpus of secondary data. Essentially, by using secondary data, you are incorporating into your research design the decisions that were made by the original researcher in their research design. Consequently, it is important to understand the research question that motivated those earlier decisions to understand more fully the data that you will be using.

In sum, a research design should not only describe how and from which sources data will be collected, it should explain how it is expected that the data will be analyzed. In constructing the research design, it is necessary to encompass both data collection and analysis.

Does the Research Design Work? Pilot and Change

A research design may look good on paper, but it is important to make sure it works in practice. Initial research designs are typically riddled with optimistic assumptions about the way sources of data will behave, about the efficacy of data elicitation, and about the validity and reliability of the data collected. Testing whether optimism is justified is essential, and "piloting" the design is necessary. Piloting entails checking whether every element of the design actually does what it is supposed to do. For instance, do desired sources agree to participate and provide data? Are any controls and manipulations of the context of data collection working reliably? Can data be systematically and comprehensively recorded in an analyzable form? Are the data collected of high enough quality – consistent with what will be needed in the main study? A pilot study is usually scaled down in the number of sources used (e.g., individuals in the sample) but should examine the full range of steps in the research process and every type of data to be elicited.

Such piloting will reveal at least some of the flaws in a design. It will point to where the design needs to be modified, but it probably will not say exactly what these modifications should be. Indeed, it is possible that several phases of piloting will be necessary before all the weaknesses in a design are eliminated. It is important not to pilot and then change a design without testing the effect of the change that has been introduced. There is always the temptation to cut corners on iterative piloting. These are temptations best to resist even at the cost of time, effort, or funding. Treated with respect, piloting will improve the design. Of course, sometimes piloting can reveal that an entire design is built around a fundamentally flawed assumption. For example, it may prove that the relevant data are simply non-existent (e.g., previously destroyed) or totally inaccessible (e.g., legally curtailed). The research question may still be potentially answerable, but the research design is thrown back to the drawing board.

Challenges to the practical viability of a research design are inevitable. It means that the process of choosing or constructing a design is one of trial and sometimes error. Admitting the possible value of changing a design and being willing to be flexible in the pursuit of an optimum solution may be difficult in practice but may pay off in the quality of the research outcomes. Being capable of abandoning or rejecting a research design that has flaws is an important research skill.

The Practicalities That Constrain Choice

Clearly, many things influence decisions related to planning a study. Some factors are more practical than others that have been considered above. They are nevertheless worth mentioning for completeness. The first is time. Do you actually have the time available to conduct the research? This boils down to whether you and/or your research team can commit to devoting the time necessary to execute the research design. Two research designs that could each be used to address a question may require quite different levels of time commitment on your part. The time that you have available in reality may be the crucial factor in deciding which design is the one to go for.

Similar to time is the consideration of timing. Is there a critical period of time when the research must be undertaken? Research questions often depend on the occurrence of real-world events, and the researcher has to shape the data collection to fit into the timing of that occurrence. Some targets for research are predictable (e.g., the anniversary of some historical event). Other research targets are very hard to predict or may be very unpredictable (e.g., a terrorist attack). Uncertainties surrounding the emergence of the critical event may make some types of research designs untenable or unattractive. For instance, using a structured experimental design may be impractical unless the researcher can predict when a particular event will occur so that measurements can be taken immediately before and after the event. It is useful to remember that research designs each have their own assumptions about their timeline relative to the researched object.

The next consideration is financial. Research designs vary in their costs (e.g., staffing, equipment, payment to respondents, or for access to databases). This is obvious, but your budget will constrain the choices available to you. Often, the researcher's task is to conduct a series of cost–benefit analyses when choosing between research designs. If obliged to do this, it is useful to have detailed in advance what each alternative research design can offer and to have determined which things are absolutely necessary for the research project to be worthwhile.

Another consideration that affects many in the social and behavioral sciences is the extent to which a study may be eligible and competitive for external funding. Of course, the research priorities set by funding agencies (whether public or private) massively influence which research questions attract researcher interest. These organizations also have an influence upon the research designs that are used. They do this directly by calling for proposals that use particular types of design. This appears to happen most frequently when the funder has sought expert advice within the research community, and some consensus emerges to suggest that the optimum approach to the research question relies upon a single or narrow range of research designs. Funders also indirectly influence the decisions researchers make by the patterns of their rejection of bids that do not incorporate the preferred design features. For instance, the increased international focus of funders in the first decades of the twenty-first century on "grand challenges" associated with societal crises (e.g., climate change or global pandemics) and on the interdisciplinary responses they necessitate has promoted the use of complex research designs that import and integrate methods from many different discipline bases (Rylance, 2015). The power of research funders to determine the choices that are made about research designs should not be underestimated.

It is also important to be sure that you and the research team have the skills needed to deploy the methods of data collection and analyses required by the research design you choose. Realizing halfway through a study that you are collecting some data that you do not know how to analyze is annoying. Becoming aware, as the study is under way, that you lack an essential skill needed to execute the design (e.g., not having the knowledge to operate a sophisticated recording or measurement instrument) can be more than annoying. It may undermine the entire study because it results in you losing vital data that may not be open to recapture. It is good practice to do a skills audit as a matter of routine before committing to using a particular research design.

The choice of research design can also be swayed by public opinion and awareness. Some research formats elicit opprobrium, and others induce incomprehension. The media and social media channels available for opinions about research are now prolific. The issue has become increasingly important as social and behavioral researchers have focused more upon attempting to do work that has impact and can influence socio-economic and political agendas and policy (Archer & Lloyd, 2002). Public understanding and support for the research enterprise have grown in significance, in part because it influences which research funding agencies are willing to support. Researchers should be alert to the way their research will be received by the public. They have an obligation both to abide by the ethical codes of research practice that their professions establish and also consider the societal relevance and intelligibility of their methods. It would be strange if this did not affect their choice of research design. For instance, the shift in the psychological sciences toward more co-production of research designs with the people who are participants in a study (Bell & Pahl, 2018) signals a shift in the nature of the relationship between the researcher and the researched.

While public opinion is important, many researchers might consider the opinion of other members of their own scientific community as important or more important. The choice of research designs and methods is heavily influenced by the prevailing orthodoxies within a discipline. This influence travels through many channels – the preferences of journal publishers, reviewers and editors, the hierarchies that determine professional advancement and recognition, as well as the support offered by funding organizations. These contextual factors affect both the choice of research design and the initial choice of the research question. The researcher is a part of a social system that will limit freedom of choices made in the research process.

Ultimately, these decisions relate to your own habits and reputation. Researchers are human too. Across their career, some people lose their flexibility when it comes to choosing a research design. They use the methods and forms of analysis that they know and for which they are known within their discipline. Habit and their established personal reputation channel their research choices. You might wish to do a little bit of research on this yourself. Choose at random three leading figures in your own discipline and check which research designs they have used across their careers. How does the pattern develop as they achieved greater seniority? Of course, it could just be that you stand a better chance of becoming a leading figure if you become associated with a particular paradigm. Also, it could be that some theoretical models with which an individual was engaged are tested most effectively using particular research designs. In this case, it would not merely be habit that drove consistent use of a design. What research design would you prefer to use to examine whether leading researchers change the range of research designs they use across their career? If there were changes, how would you determine why change occurs? Of course, there may not be changes. Some people find a research design that suits them very early in their career and stick with it through thick and thin.

A Conclusion: It Pays to Invest the Time in Optimizing the Research Design

It is actually quite hard to stand back and look at the full array of designs that you could use to address a research question. It can be a daunting task to do the cost–benefit analysis that should be the basis for your decision. However, it can also be a rewarding exercise. Thinking seriously about different designs will enable you to see your research question in different ways. You may even see new components of your research question. One message of this chapter is that constructing a research design can be an intellectual adventure. There is also another message – a sort of subtext. The latitude afforded the researcher in making choices about a research design is limited. The objective of this chapter is to offer a checklist of the things to be considered during the construction of a research design. Taking the time to understand the constraints that exist is time well spent. It can make the difference between doing research that is dismissed as trivial, irrelevant, or erroneous and doing research that is lauded as a significant contribution. As Martin Luther King Jr. said "You must learn, research, and be so passionate about new ways and methods of doing things to remain relevant." Good research design is the foundation for real research relevance.

References

Archer, J. & Lloyd, B. (2002). *Sex and Gender*. Cambridge University Press.

Bakker, J. I. (2010). Interpretivism. *Encyclopaedia of Case Study Research*, *1*, 486–493.

Bell, D. M. & Pahl, K. (2018). Co-production: Towards a utopian approach. *International Journal of Social Research Methodology*, *21*(1), 105–117.

Blaikie, N. & Priest, J. (2019). *Designing Social Research: The Logic of Anticipation*. John Wiley & Sons.

Clarke, V. & Braun, V. (2014). Thematic analysis. In t. Teo (ed.). *Encyclopedia of Critical Psychology* (pp. 1947–1952). Springer.

Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences*. Academic Press.

Breakwell, G. M. (2014). *The Psychology of Risk*. Cambridge University Press.

Breakwell, G. M. & Fife-Schaw, C. R. (1994). Using longitudinal cohort sequential designs to study changes in sexual behaviour. In M. Boulton (ed.), *Challenge and Innovation: Methodological Advances in Social Research on HIV/AIDS* (pp. 25–38). Taylor & Francis.

Breakwell, G. M., Wright, D. B., & Barnett, J. (2020) Research questions, design, strategy and choice of methods. In G. M. Breakwell, D. B. Wright, & J. Barnett (eds.), *Research Methods in Psychology*, 5th ed. (pp. 1–30). Sage.

Doise, W. & Valentim, J. P. (2015). Levels of analysis in social psychology. In J. D. Wright (ed.), *International Encyclopedia of the Social & Behavioural Sciences*, 2nd ed. (pp. 899–903). Elsevier.

Fife-Schaw, C. (2020). Quasi-experimental designs (including observational methods). In G. M. Breakwell, D. B. Wright, & J. Barnett (eds.), *Research Methods in Psychology*, 5th ed. (pp. 161–180). Sage,

Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin*, *87*(3), 564–567.

Hole, G. (2020). Experimental design. In G. M. Breakwell, D. B. Wright, & J. Barnett (eds.), *Research Methods in Psychology*, 5th ed. (pp. 182–216). Sage.

Jaspal, R. (2020). Content analysis, thematic analysis and discourse analysis. In G. M. Breakwell, D. B. Wright, & J. Barnett (eds.), *Research Methods in Psychology*, 5th ed. (pp. 285–312). Sage.

Nosek, B. A. & Errington, T. M. (2020). What is replication? *PLoS Biol 18*(3), e3000691. doi: https://doi.org/10.1371/journal.pbio.3000691.

Ponterotto, J. G. (2005). Qualitative research in counselling psychology: A primer on research paradigms and philosophy of science. *Journal of Counselling Psychology*, *52*(2), 126.

Rylance, R. (2015) Grant giving: Global funders to focus on interdisciplinarity. *Nature News*, *525*(7569), 313.

Semino, E. (2017). Corpus linguistics and metaphor. In B. Dancygier (ed.), *The Cambridge Handbook of Cognitive Linguistics* (pp. 463–476). Cambridge University Press.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103, 677–680.

von Braun, W. & White, H. J. (1953). *The Mars Project*. University of Illinois Press.

Yardley, L. (2000). Dilemmas in qualitative health research. *Psychology and Health*, *15*(2), 215–228.

6 Building the Study

Martin Schnuerch and Edgar Erdfelder

Abstract

This chapter discusses the key elements involved when building a study. Planning empirical studies presupposes a decision about whether the major goal of the study is confirmatory (i.e., tests of hypotheses) or exploratory in nature (i.e., development of hypotheses or estimation of effects). Focusing on confirmatory studies, we discuss problems involved in obtaining an appropriate sample, controlling internal and external validity when designing the study, and selecting statistical hypotheses that mirror the substantive hypotheses of interest. Building a study additionally involves decisions about the to-be-employed statistical test strategy, the sample size required by this strategy to render the study informative, and the most efficient way to achieve this so that study costs are minimized without compromising the validity of inferences. Finally, we point to the many advantages of study preregistration before data collection begins.

Keywords: Validity of Studies, Hypothesis Testing, Estimation, Sampling Strategies, Power Analysis, Preregistration

The Value of Preparation

> *Give me six hours to chop down a tree and I will spend the first four sharpening the axe.*
>
> Abraham Lincoln (allegedly)

It is not documented whether Abraham Lincoln ever cut down a tree, let alone in six hours. Maybe he never even said those words himself. Nevertheless, the quote reflects an important aspect of successful project management: *preparation*. Preparation is not only essential if you are a lumberjack (or president of the United States, for that matter), but it is also a crucial step in the research process. Before we start collecting and looking at data, we need to make sure that the data allow for a meaningful conclusion about our research question. In this chapter, we focus on the preparation process of social and behavioral scientists: How can we build the study such that its results allow for valid inferences about our research question?

As empirical scientists, we advance our knowledge by critically testing how the predictions derived from a theory fare against experience (the Greek word *empeiría* means "experience"). Put simply, if what we observe is not in line with the theory's predictions, we may take it as evidence against the theory. Not everything we observe, however, allows for inferences about the theory. Consider the following example: According to the levels-of-processing theory (Craik & Lockhart, 1972), deeper processing of information leads to better memory for that information.

https://doi.org/10.1017/9781009010054.007 Published online by Cambridge University Press

One prediction we can derive from this theory is that if people learn words and non-words (i.e., random letter strings), performance in a subsequent memory task should be better for words because they allow for deeper processing. However, what if we presented English words, and all the people that we tested did not know any English? Obviously, there is no reason to expect that they process words they do not know deeper than non-words. Consequently, our experiment would not actually test the levels-of-processing theory.

If the study design (i.e., the manipulation, the measured variables, characteristics of the sample, or the analysis) leaves no realistic chance for the prediction to be supported or fail, the study likely produces misleading results and wastes valuable resources. Unfortunately, reports of low replication rates indicate that several published findings in social and behavioral sciences may indeed have been misleading (Pashler & Wagenmakers, 2012).

When building a study, the goal is to ensure that the data we observe are maximally informative about our underlying research question. Thus, the very first step is to carefully specify the primary aim of the study: What is the research question we want to answer? Lin et al. (2021) identify four possible aims of an experimental study: (1) to test a hypothesis/theory, (2) to search for novel phenomena, (3) to develop theories, and (4) to advise policy makers. We can further simplify this taxonomy into a rough, but very helpful dichotomy: testing versus estimation. Both terms closely relate to what the philosopher Reichenbach (1938) called *context of justification* and *context of discovery*, respectively (Erdfelder, 1994).

Testing refers to the critical evaluation of a theory. To test a theory, we derive predictions (hypotheses) and test these in light of data. This approach is typically referred to as a *confirmatory* or *deductive* approach. It constitutes a vital step in the scientific process because theories can only be refuted and improved if we critically scrutinize their predictions and detect where they fail (Lakatos, 1978; Popper, 1968). Thus, the goal in hypothesis testing is to ensure an accurate answer to our question: Do the data support the hypothesis?

Estimation, on the other hand, refers to the assessment of some quantity. For example, instead of testing whether memory performance for words is better than for non-words, we might want to estimate *how large the difference* in memory performance is. Thus, estimation reflects an *exploratory* or *descriptive* approach. Consequently, the goal in estimation is precision: What is the numerical size of a certain quantity (e.g., an effect)?

There have been many debates over which of the two aims is superior (e.g., Cumming, 2014; Kruschke, 2013). Ultimately, however, "neither hypothesis testing nor estimation is more informative than the other; rather, they answer different questions" (Morey et al., 2014, p. 1290). We may use estimation to derive more precise predictions or to refine theories; hypothesis testing is necessary, in contrast, to critically assess a theory's predictions.

To build an informative study, we need to be clear about what question we want to answer. In this chapter, we focus on studies that aim to test a hypothesis. We caution the reader to keep in mind, however, that priorities in the specification of study parameters may be different if the research question is different.

Implementing the Study

The extent to which a study is informative (i.e., allows for conclusions about the underlying research question) defines the study's validity. Validity refers to the control of confounding extraneous variables that would compromise our interpretation of the results. Recall our example: We could not interpret a lack of a difference in memory performance between words and non-words as support against the levels-of-processing theory because there is an alternative explanation (i.e., that our manipulation did not induce different processing levels). Thus, if we do not control for confounding influences, our study does not afford a valid interpretation.

Validity also refers to the generalizability of our conclusions. In 1957, in a now classic article, Donald Campbell described two ways in which a study allows for meaningful interpretations (Campbell, 1957). The first one is *internal validity* – Does the study allow for a causal interpretation of the observed effect (i.e., group or treatment difference)? Internal validity increases with the extent of experimental control over all relevant influences. The second one is *external validity* – Can we generalize our interpretation to a different, possibly non-experimental context beyond our study? External validity increases with the extent to which the study context is representative of real-life conditions.

Obviously, an ideal study would have both high internal and external validity. Unfortunately, the two are often at odds with each other (Lin et al., 2021). While *laboratory experiments* with high internal validity often create highly controlled, artificial contexts with limited generalizability to other contexts, it is often impossible to experimentally control or manipulate confounding influences in *field studies* or *quasi-experiments* that take place in more realistic contexts with high external validity. Some have argued that controlling confounding influences and isolating effect mechanisms under experimental conditions should generally be prioritized (Falk & Heckman,2009). Others have warned about the risk of a *mutual-internal-validity problem* where theories become increasingly tied to specifics of an experimental paradigm, thus losing relevance for a reality outside the laboratory (Lin et al., 2021;Meiser, 2011).

To conclude, when designing a study, we should aim for striking the right balance between internal and external validity (Leatherdale, 2019). Importantly, such a balance also depends on the research question (Schram, 2005). While critical tests of theories should prioritize internal validity (*context of justification*), studies aiming at the exploration of novel phenomena or the development of theories require a stronger focus on external validity (*context of discovery*).

Choosing the Appropriate Sample

Typically, the observations that we collect in an experiment constitute only a subset of all the possible observations we could have made. Assume, for example, that we tested the levels-of-processing theory in an experiment with 100 English-speaking adults. Obviously, the individuals that we tested in this experiment are just one possible subset of the universe of people that we could have tested.

We call the subset of observations a *sample*, and we call the "totality of potential units for observation" a *population* (Hays, 1963, p. 192). In most cases, we are not interested in the specific sample, but we want to make general statements about the underlying population. Because we cannot measure the entire population, we collect a sample and infer from the sample what we want to know about the population. This is called *statistical inference*.

To build our study such that it can yield meaningful inferences about our population, we must first define it: To what entity should our inferences refer? For example, do we want to make a statement about all people, or just the people in a specific country or age range? Our research question defines our population. In many areas of social and behavioral sciences, research questions are concerned with humans in general. Hence, the population to which we want to infer our conclusions is all human beings.

Beside the question that we want to answer, the population about which our sample allows to draw conclusions is also affected by the sampling strategy (Hays, 1963). Imagine that we wanted to estimate the average height of citizens of the European Union. To do so, we randomly select 100 Dutch citizens and measure their height. Does our sample allow for a meaningful estimate of the average European's height? Considering the notable variance of average height across European nations, it certainly does not. If we only sample individuals from the Netherlands (who are among the tallest people in Europe; Guven & Lee, 2015), we may draw inferences about the Dutch population but not necessarily all Europeans.

The ideal case, in which our sample would be guaranteed to allow for unbiased inference about our population, would be a *representative* sample. A sample is said to be representative if it reflects the characteristics of the population (at least with respect to certain criteria of interest). To estimate the average height of Europeans, the structure of nations, gender, and age represented in our sample should reflect the structure in the population of Europeans. If the sample is not representative and our aim is to get a precise estimate of a certain quantity, our inference may be biased.

There are several sampling strategies, some of which aim at forming a representative sample. We will discuss three common strategies. In a *random sampling* strategy, every individual of the population has the same probability to be sampled. Thus, if the sample is large enough, the characteristics of the sample can be expected to reflect the population's characteristics. Hence, inference from a random sample is unbiased and allows for high generalizability to the underlying population. In practice, however, random samples are very difficult to achieve. Depending on the population, it may be very hard – or even impossible – to ensure that every individual has the same probability to be sampled. Thus, successful implementations of random sampling strategies rarely occur in behavioral research (Highhouse & Gillespie, 2009).

In *stratified sampling*, we define relevant, homogeneous subgroups (so-called *strata*) of the population. For example, with respect to the average height of Europeans, these subgroups may be defined by the combination of nation, gender, and age group. Samples are then drawn from these subgroups. Importantly, while the samples from each subgroup are drawn at random, the proportion of draws from each group reflects the proportion of this group in the population. Thus, with this strategy, we may arrive at a representative sample, at least with respect to the considered criteria.

In contrast to these two so-called *probability sampling* methods, social and behavioral scientists often rely on a different, *non-probability sampling* strategy. In *convenience sampling*, the sample is drawn from a certain subset of the population that is conveniently available. Consider a political scientist interested in election preferences of voters in a particular country. To estimate the proportions of party preferences, the scientist may survey people about their election preferences in a shopping mall. The individuals who are surveyed may be selected at random, but the sample is restricted to people who shop in this particular mall. Hence, the sample can make no claims of being representative of the population (i.e., all voters in that country).

Social and behavioral researchers often rely on convenience samples, typically undergraduate students or, increasingly, users of crowdsourcing platforms such as Amazon Mechanical Turk (Highhouse & Gillespie, 2009). Does this widespread reliance on convenience samples pose a threat to the validity of behavioral research? Unfortunately, the answer is not that simple. Plainly labeling convenience sampling as a bad strategy would be throwing out the baby with the bathwater. Instead, one should carefully consider the advantages and disadvantages of a sampling strategy in the context of one's particular research question (Landers & Behrend, 2015). Convenience samples have a number of advantages that, in certain situations, may outweigh possible limitations (Highhouse & Gillespie, 2009). For example, convenience sampling is a cost-efficient strategy to gather data. Moreover, the samples are often more homogeneous and, thus, less noisy than samples gathered with other (probability) sampling strategies (see Jager et al., 2017).

As always, the choice of an appropriate sample cannot be made without reference to the underlying research question. If the aim is to describe a well-defined population as accurately as possible (context of discovery), it may be important to employ a sampling strategy that ensures a sample that reflects the relevant characteristics of the population. If, however, the aim is to test a hypothesis that refers to all individuals (context of justification), it is important that the sample "be relevant – this is, that it fit within predefined population/universe boundaries" (Sackett & Larson Jr., 1990, p. 435). In other words, unless the subpopulation from which we draw a sample differs systematically from the target population on dimensions relevant to the hypothesis, any subpopulation is appropriate to test this hypothesis. Thus, in confirmatory studies aiming at testing the predictions derived from a theory, convenience samples typically do not pose a threat to the study's validity (e.g., Bredenkamp, 1980). Nevertheless, careful considerations of possible constraints of the specific convenience sample at hand are always warranted (Landers & Behrend, 2015). See Chapter 9 in the volume for a larger discussion of issues related to sampling.

Choosing Appropriate Statistical Hypotheses

After we have chosen the appropriate subpopulation to sample from, we want to test whether a hypothesis (i.e., a prediction derived from a theory) holds in this subpopulation. “A prediction is a statement of what a theory does and does not allow” (Roberts & Pashler, 2000, p. 359). Thus, to test a theory we need a means to quantify whether what we observe is allowed by the theory (thus, supporting it) or not (thus, refuting it).

In practice, this is easier said than done because theories in the social and behavioral sciences are typically verbal and do not provide an explicit link to observed data (Erdfelder & Bredenkamp, 1994). Recall our example on testing the levels-of-processing theory. What is “deep processing”? What is “better memory”? We *operationalize* processing depth by means of meaningfulness (words vs. non-words) and measure memory by means of the number of recalled items. However, what does our theory say about the number of recalled items? Is our theory falsified as soon as we observe a single person that recalls more non-words than words?

Our verbal theory does not afford a precise prediction for the number of recalled items. Moreover, data are error-prone. Specifically, the number of recalled items is not a perfect measure of memory performance; it is always affected by noise (i.e., by random variation due to unsystematic influences). Thus, to test the theory’s prediction, we need to formalize it in a way that links substantive theory to the data. Such a link is provided by *statistical models* (Rouder et al., 2016).

Statistical models describe the probabilistic structure underlying observed data, taking variability and noise into account. For example, we could describe the number of recalled items as a random draw from a normal distribution. Depending on the location (mean) and the spread (variance) of this distribution, some data are more likely to be observed than other data. Importantly, these parameters are informed by our hypotheses (i.e., by the substantive prediction derived from our verbal theory). This is a *statistical hypothesis* – constraints on the parameters of the statistical model implied by the substantive prediction of our theory. Thus, by means of a statistical hypothesis, we can formalize the verbal theory, make it precise, and allow for a quantitative test against data (Vanpaemel & Lee, 2012).

For our example, we may specify two normal distributions with common variance σ^2 and means μ_w and μ_n for the number of recalled words and non-words, respectively. Our theory is about the systematic difference between memory performance for words and non-words. Thus, we can express it in statistical terms as a constraint on the means of the specified distributions (i.e., $\mu_w - \mu_n > 0$). We call this systematic mean difference an *effect*, namely, the effect of processing depth on memory performance. The standardized quantification of this effect is known as the *effect size*. In this example, we could express the effect size in terms of Cohen’s $d = (\mu_w - \mu_n) / \sigma$.

To test a theory, we must specify our expectation under two possible states of the world: if the theory were true and if it were false (Morey et al., 2014). Thus, the test of a hypothesis is, in fact, a test of competing hypotheses – the *null hypothesis* and the *alternative hypothesis*. Conventionally, the null hypothesis represents the absence of an effect. In our example, the null hypothesis posits that there is no systematic difference between the number of recalled words and non-words. That is, it represents the state of the world if the underlying levels-of-processing theory was false. In statistical terms, this null hypothesis states that the effect size is zero (i.e., $d = 0$). Given this constraint on the parameters of our statistical model, we can specify precisely which data are likely and which are less likely to be observed. This allows us to evaluate how poorly the data agree with the null hypothesis – the basis of the commonly used *null-hypothesis significance testing* (NHST).

Logically, the complement of the above-stated null hypothesis would be $d \neq 0$ – the implicitly assumed alternative hypothesis in NHST. Unfortunately, this alternative hypothesis does not allow for a probability statement about the data. Without specifying which data we expect under the alternative hypothesis, we cannot test the underlying theory (Morey et al., 2014). Thus, we must specify substantively motivated constraints on model parameters under both the statistical null and alternative hypothesis. The specification of testable statistical constraints is often one of the greatest challenges when building a study.

There are three common ways to specify precise statistical hypotheses. The ideal case is that the theory has been formalized to the extent that it makes quantitative predictions that can directly be translated to a statistical hypothesis. For example, the total-time hypothesis (TTH) of verbal learning states that performance in multi-trial learning depends on the total learning time, not on the time allowed per learning trial (Cooper & Pantle, 1967). We could test this by comparing two experimental groups who study the same item list, but with, say, $t = 10$ versus $t = 20$ seconds allowed per learning trial. The observed dependent variable is L – the number of learning trials required until perfect mastery of all items. If μ_{10} and μ_{20} represent the means of L in the 10- and the 20-seconds study condition, respectively, the TTH obviously predicts $\mu_{10} \cdot 10 = \mu_{20} \cdot 20$ or, by implication, $\mu_{10} = 2 \cdot \mu_{20}$. A straightforward way to test this precise alternative hypothesis is by means of the transformed dependent variable $L^* = L \cdot t$ – the total learning time per participant – for which the TTH implies equal means in both experimental conditions. Thus, this example shows how precise theoretical predictions almost always can be translated to statistical null hypotheses, either directly or indirectly, after suitable transformation of dependent or independent variables as implied by the hypothesis of interest (Erdfelder & Bredenkamp, 1994). Moreover, it also demonstrates that null hypotheses don't necessarily represent the absence of empirical regularities; quite the contrary, they may represent a precisely formalized expectation of an invariance relation across conditions (Rouder et al., 2009).

A different approach is to specify hypotheses based on information that we gather from the literature. If previous studies or meta-analyses on the effect of interest have been conducted, the results may give an informed estimate of the effect size that can be expected under the alternative hypothesis. Due to systematic distortions, such as publication bias, however, single-study estimates and even meta-analytic estimates may overestimate the effect size. Therefore, when drawing on previous studies, it may be prudent to use meta-analytic estimation methods that account for possible publication bias (e.g., Ulrich et al., 2018) or to rely on a conservative lower-bound estimate (Perugini et al., 2014).

The most common strategy is to specify the minimum effect size that would be of practical interest (see Cohen, 1988). The rationale underlying this strategy is that a study designed to be informative with respect to the specified effect is also informative with respect to larger effects. Thus, if we specify the smallest effect size of interest and build the study accordingly, we can ensure that it is informative for any practically relevant effect size. For our levels-of-processing example, we may specify as the smallest effect size of interest a Cohen's $d = 0.20$. Conventionally, this is considered a small effect (Cohen, 1988). It is important to keep in mind, however, that these conventions may have different meanings across disciplines and test procedures (Faul et al., 2007).

So far, we have focused on specifying statistical hypotheses by expressing expectations about particular effect sizes. Some authors have argued, however, that the specification of a single effect size is too restrictive and often unjustifiable (e.g., Rouder et al., 2016). Instead, these authors suggest that we express uncertainty by means of so-called *prior probability distributions*. These distributions indicate which effect sizes we deem likely and unlikely by putting varying probability mass on the respective values. To evaluate the probability of observed data, we can then integrate across the prior distribution, thus calculating a weighted average probability under all possible effect sizes. This approach is known as the *Bayesian* approach (discussed in more detail in Chapter 23 of this volume) and has recently become increasingly popular among social and behavioral scientists.

Whatever strategy one chooses to formulate precise statistical hypotheses, it is important to keep in mind that the challenge is not merely a statistical one. Statistical hypotheses are instantiations of substantive hypotheses; they put our knowledge and our theoretical assumptions about the processes under scrutiny into a formal, mathematical structure (Vanpaemel, 2010). Thus, the specification of statistical models and hypotheses is a crucial step in building a study, and it is primarily a substantive challenge (Rouder et al., 2022).

Choosing the Test Procedure

By specifying a statistical null and alternative hypothesis, we have formally expressed what we expect if the theory were true versus if it were false. To test our theory in light of data, we further need a principled method to decide whether the data that we observe support one or the other hypothesis (Morey et al., 2014). We call this principled method the *test procedure*, and like the formulation of statistical hypotheses, the choice of a certain test procedure is crucial in the process of building a study.

In the past, textbooks and curricula in the social and behavioral sciences have often presented NHST not as one instance but rather the archetype of a test procedure (Gigerenzer, 2004). For as long as it has been around, however, it has also been criticized (e.g., Bakan, 1966; Bredenkamp, 1972; Gigerenzer, 1993; Wagenmakers, 2007). The main point of critique is that NHST is an imbalanced procedure – Researchers specify the null hypothesis and calculate the *p*-value (i.e., the probability to obtain the observed or a more extreme test statistic under the null hypothesis). The p-value is then compared with the pre-specified α-level that denotes the probability to reject the null hypothesis when it is true (Type I error). If the p-value falls below α, the null hypothesis is rejected. If it is larger, however, the procedure does not allow for a conclusive decision. Thus, we can either reject or fail to reject the null hypothesis, but we can never accept it.

Fortunately, statistical inference is not just a single tool but rather an *adaptive toolbox* with different tools suited for different aims and problems (Gigerenzer, 2004). It is the individual researcher's responsibility to carefully choose that procedure from the toolbox which is suited for her individual situation (Lakens, 2021). And this choice also affects the design of the study (e.g., the sample size). There is no blueprint for an informative study; to maximize informativeness, the design must be optimized with respect to the aim of the study and the accordingly chosen test procedure (Heck & Erdfelder, 2019).

We will discuss two procedures in the following. In the Neyman–Pearson approach, the hypothesis test constitutes a decision-making procedure. While NHST only specifies a null hypothesis without an explicit alternative, Neyman and Pearson acknowledged the importance of specifying both the null and the alternative hypothesis in the decision-making process. By additionally considering the alternative, we can quantify the probability to falsely retain the null hypothesis (Type II error). Thus, the Neyman–Pearson approach provides a method to arrive at a statistical decision while controlling both the probabilities of a Type I and Type II error.

The Neyman–Pearson procedure is particularly relevant in situations that compel researchers to act in a certain way (e.g., to implement a new therapy, to abandon a line of research based on the outcome of a pilot study, or simply to make the claim that a hypothesis has been supported or refuted; Lakens, 2021). Some authors have argued, however, that researchers are more interested in quantifying the *statistical evidence* that the data provide for the hypotheses (e.g.,Bakan, 1966; Edwards et al., 1963; Wagenmakers, 2007). This perspective is at the heart of *Bayesian hypothesis testing*.

In Bayesian testing, it is assumed that we hold prior beliefs about how plausible the hypotheses are. The aim of a hypothesis test is then to update these prior beliefs by looking at data. The extent to which beliefs are updated is the extent to which the data provide evidence for one hypothesis over the other. Thus, instead of making a decision to accept or reject a hypothesis, we quantify the statistical evidence in the data.

For quite some time, there has been a dispute between proponents of the Bayesian procedure and those of the Neyman–Pearson procedure over which is better suited to address scientific research questions. Instead of arguing over which one is superior, however, we should acknowledge that both have strengths and weaknesses and carefully choose the procedure best suited to our research question: Do we want to assess statistical evidence and update subjective beliefs? Then we may choose a Bayesian procedure. Or do we want to take specific actions based upon accepting or rejecting our hypotheses? Then it is important to choose a procedure that controls the probabilities to commit a decision error. Once we have chosen a procedure, we can optimize the design to provide for a maximally informative study (Heck & Erdfelder, 2019).

Determining the Sample Size

Intuitively, sample size determination could be guided by a simple heuristic: “The more, the better”. This is not necessarily true, however, because collecting observations requires resources (e.g., time, money, and laboratory facilities). Scientific research is an expensive endeavor that devours a great deal of public resources (Miller & Ulrich, 2020). Therefore, scientists have an obligation to put these resources to good use. At a certain point, the gain in information by sampling additional observations no longer justifies the costs; therefore, the sample size should be carefully justified to avoid squandering valuable resources (see Lakens, 2022, for a review).

A simple way to determine the sample size is by reference to some heuristic or previous study. As elaborated above, however, the informativeness of a study depends critically on the specifics of the study. A reference to a previous study only makes sense if that study’s characteristics and justification for the sample size also apply to the current study. If not, a recourse to previous studies or simple heuristics is ill-advised (Lakens, 2022).

A convincing sample size justification is based on an a priori power analysis (Cohen, 1988). The *power* of a statistical test is the complement of its Type II error probability β (i.e., $1 - \beta$). Thus, it denotes the probability to reject the null hypothesis, given that the alternative hypothesis is true. Importantly, for a given effect size (i.e., for given statistical null and alternative hypotheses), the statistical power of a test is a function of the Type I error probability and the sample size. Consequently, with an a priori analysis, we can optimize the sample size for the effect size of interest such that the test has the desired error probabilities and power, respectively.

The error probabilities of a statistical test reflect the fairness (i.e., the probability of confirming true hypotheses) and the severity (i.e., the probability of rejecting false hypotheses) with which the hypotheses have been tested (Erdfelder & Bredenkamp, 1994; Mayo, 2018). Depending on the consequences of false confirmations or rejections, we may want our test to be more or less fair or severe. For example, if we tested the efficacy of a new treatment, which error is more harmful: Wrongfully accepting the hypothesis that it is effective or wrongfully maintaining that it is not? The challenge to choose “the right” error probabilities is certainly a tough one. Researchers often rely on widespread conventions such as $\alpha = 0.05$ and $\beta = 0.20$ as recommended by Cohen (1988). It should be obvious, however, that these conventions cannot reflect the individual severity of decision errors in every study. Therefore, the burden of justifying the error probabilities rests again on the individual researcher (Lakens et al., 2018).

Once the choice is made, the determination of the required sample size is straightforward. There are two approaches to calculate the sample size for given hypotheses and error probabilities. The first one is by analytical calculation. If the distribution of the test statistic under both hypotheses is known, we can solve it analytically for the required sample size. The advantage of an analytical solution is that it is fast and exact (given available software that we address below). Analytical power analyses can be used in the context of many common test procedures, including ANOVA, ANCOVA, *t*-tests, *z*-tests, χ^2-tests, and multiple linear regression.

The second approach is based on Monte Carlo simulations and can be used if the sampling distribution of the test statistic is unknown or cannot be solved analytically. The principle is quite simple: We generate random samples (e.g., 10,000) with a specific sample size from the statistical model representing the alternative hypothesis and calculate the test statistic for each simulated sample. By summarizing the simulated distribution, we can estimate the power (i.e., the proportion of test statistics exceeding the critical value). If the power is lower than desired, we repeat these steps with larger sample sizes and choose the one that satisfies the required error probabilities.

Simulation-based power analyses are straightforward to implement. They allow for a simple and general solution to determine the required sample size for a desired level of statistical power. However, depending on the complexity of the research design and the test procedure, they may require extensive time and computational power. Moreover, the results are not exact and may cause problems especially when small error probabilities are of interest.

Considering the importance of conducting a power analysis, it is surprising that the issue has been routinely ignored (Brysbaert & Stevens, 2018). Fortunately, there are nowadays many available software packages for power analysis. Table 6.1 provides an overview of free software tools that support power analysis for a wide range of statistical procedures. Note that this is just a selection and not an exhaustive list of available tools. Additionally, there are many accessible tutorial papers on power analysis both on simple experimental designs such as *t*-tests and ANOVA (Brysbaert, 2019; Perugini et al., 2018) and on more complex designs such as generalized linear models and mixed-effects models (Brysbaert & Stevens, 2018; Kumle et al., 2021).

In the following, we illustrate the power analysis for a simple design in one of the most widely used software tools –G*Power. Recall our example on testing the levels-of-processing theory. In the preceding sections, we formulated the statistical model for the hypothetical experiment as well as the statistical hypotheses. Based on our specifications, we want to determine the sample size such that it allows for an informative result. As desired error probabilities, we specify $\alpha = \beta = 0.05$, reflecting that neither error is considered less consequential than the other. Figure 6.1 shows a screenshot of G*Power with all relevant input and output parameters.

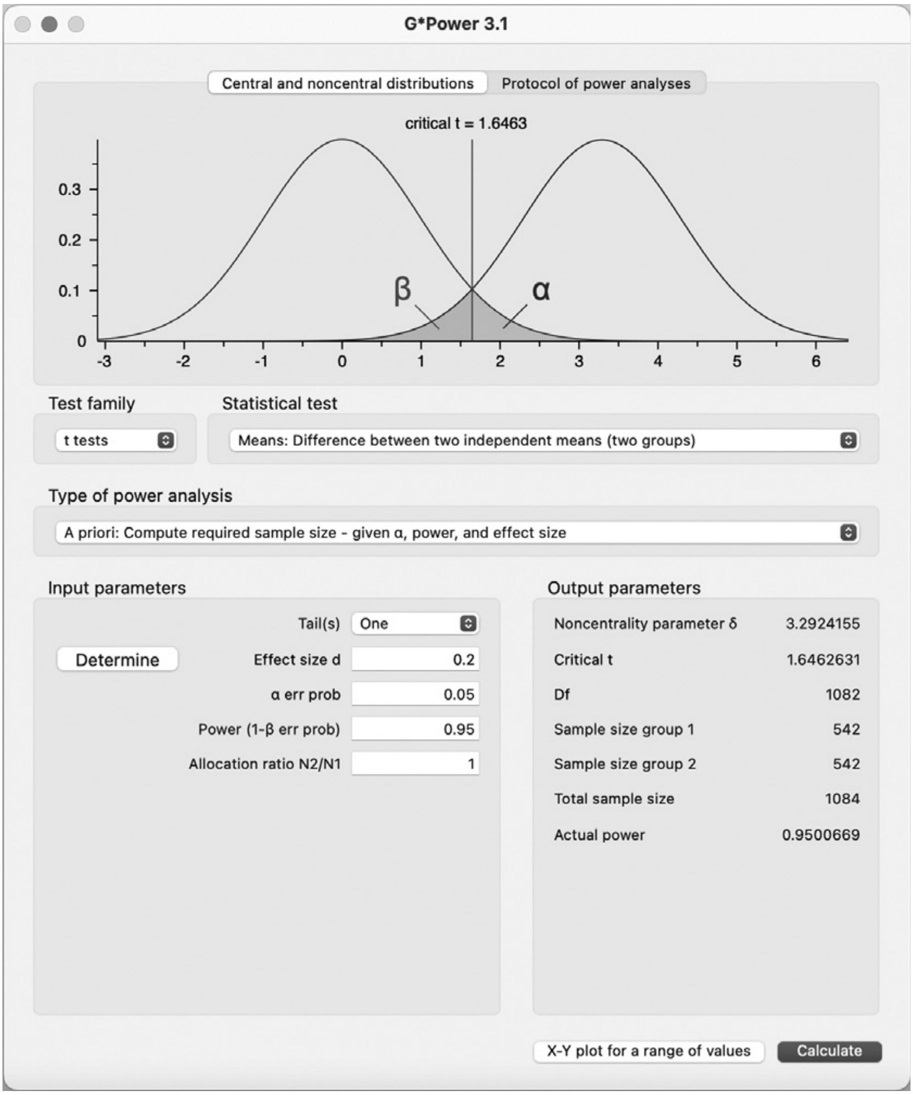
In the first step, we select the appropriate test family and test procedure. We plan to conduct a *t*-test on the difference between two independent means. As the type of power analysis, we choose a priori to indicate that we want to compute the required sample size (we will address other types of power analysis in the next section). Finally, we define the parameters of the test procedure (i.e., the direction of the test – *one-tailed* since we expect an effect in a certain direction), the expected effect size ($d = 0.20$), the

Table 6.1 *Overview of eight freely available power/design analysis tools*

Program	Type of software	Supported test procedures (examples)	Analysis approach	References
BFDA	R package and Shiny app	Bayesian analysis: *t*-tests, AB-tests, correlations	Simulation-based	(Schönbrodt & Stefan, 2019)
G*Power	Stand-alone	Tests of proportions (exact and based on normal approximation), correlations, *F*-tests (e.g., ANOVA, ANCOVA, MANOVA, and multiple regression), *t*-tests, χ^2-tests, logistic, and Poisson regression	Analytical	(Faul et al., 2007)
MorePower	Stand-alone	Tests of proportions (based on normal approximation), correlations, ANOVA, and *t*-tests	Analytical	(Campbell & Thompson, 2012)
pwr	R package	Tests of proportions (based on normal approximation), correlations, *F*-tests, *t*-tests, and χ^2-tests	Analytical	(Champely, 2020)
pwr2ppl	R package	ANOVA, correlations, *t*-tests, multiple regression, mediation analysis, and logistic regression	Analytical	(Aberson,2019)
SIMR	R package	Generalized linear mixed models	Simulation-based	(Green & MacLeod, 2016)
SSDbain	R package	Bayesian analysis: *t*-tests, Welch's tests	Simulation-based	(Fu et al., 2021)
superpower	R package and Shiny app	ANOVA (including up to three within or between factors)	Simulation-based	(Lakens & Caldwell, 2021)

Type I error probability ($\alpha = 0.05$), the power ($1 - \beta = 0.95$), and the ratio of sample sizes within each group ($n_1/\mathrm{n}_2 = 1$, indicating a balanced design). With these parameters, G*Power calculates a total sample size of 542 + 542 = 1,084 participants.

Power is a central concept in statistical decision-making. However, a priori analyses to design an informative study are not limited to power calculations (Heck & Erdfelder, 2019). Also in the context of Bayesian procedures, defining the sample size such that it maximizes the probability of obtaining an informative result is important. The general term for this is *design analysis* or *design calculation* (Gelman & Carlin, 2014). Simulation-based design analyses for Bayesian procedures have been implemented, for example, in the R packages BFDA (Schönbrodt & Stefan, 2019) and SSDbain (Fu et al., 2021).


G*Power 3.1
Central and noncentral distributions
Protocol of power analyses
critical t = 1.6463
0.3
0.2
0.1
0
-3
-2
-1
0
1
2
3
4
5
6
β
α
Determine
X-Y plot for a range of values
Calculate

Test family: t tests

Statistical test: Means: Difference between two independent means (two groups)

Type of power analysis: A priori: Compute required sample size - given α, power, and effect size

Input parameters

Tail(s): One
Effect size d: 0.2
α err prob: 0.05
Power (1-β err prob): 0.95
Allocation ratio N2/N1: 1

Output parameters

Noncentrality parameter δ: 3.2924155
Critical t: 1.6462631
Df: 1082
Sample size group 1: 542
Sample size group 2: 542
Total sample size: 1084
Actual power: 0.9500669

Figure 6.1 *Screenshot of a power analysis for a two-groups* t*-test in G*Power.*

Dealing with Limited Resources

Ideally, sample sizes would always be determined by well-justified design analyses. In practice, however, researchers face several constraints such as limited time and financial resources or limited access to participants.

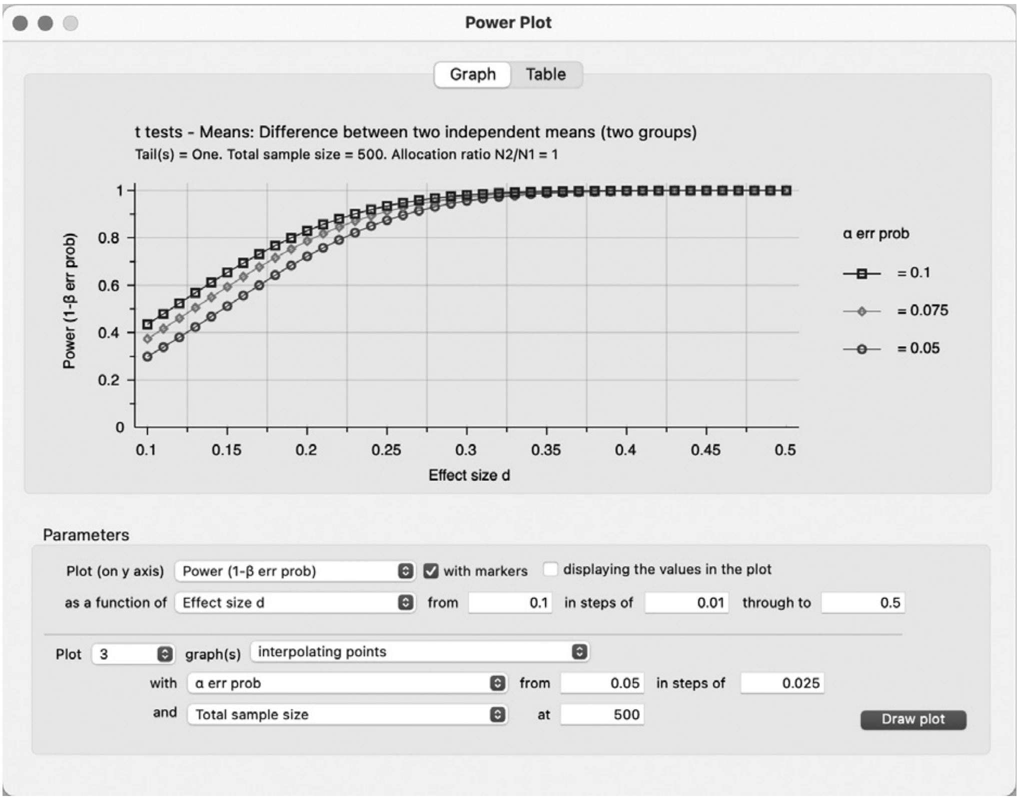
To illustrate, in our example, we specified a minimum relevant effect size of $d = 0.20$ and $\alpha = 0.05$. Assume that, due to limited resources, we were not able to sample 1,084 participants. Instead, we could only sample 250 observations per group (i.e., 500 observations in total). What are the consequences of this notably smaller sample size? This question can be answered with a so-called *post hoc power analysis*. In this analysis, we calculate the chance of our test to detect a certain effect in the population given the sample size and the other parameters of our test procedure. For our example, G*Power calculates a power to detect an effect of size $d = 0.20$ with $\alpha = 0.05$ and $N = 500$ of $1 - \beta = 0.722$.

Instead of calculating the resulting power for a certain population effect size, we can also analyze a given sample with respect to the population effect size that would result in a certain level of statistical power. This is called a *sensitivity analysis*. For example, if we could only sample 500 observations and test our hypotheses with $\alpha = 0.05$, we would require a population effect size of at least $d = 0.29$ for our test to have a chance of $1 - \beta = 0.95$ to detect it. If this is a reasonable effect size to expect, the analysis shows that we have an informative study even when the sample size is limited. If not, however, it indicates that we either need a stronger manipulation (thus increasing the effect size we can expect) or we are running the risk of conducting an underpowered study.

The best way to communicate a sensitivity analysis is by plotting the power curve (i.e., the power of the test as a function of the population effect size). In G*Power, we can create this graph by clicking the "X–Y plot for a range of values" button. Figure 6.2 displays the plot for our example with $N = 500$ observations. The plot is helpful not only for researchers who plan a study with limited resources but also for other researchers when communicating the informativeness of the study across a reasonable range of possible population scenarios (Lakens, 2022).

Post hoc and sensitivity analyses illustrate that, with a given sample size and a standard α, the power may be quite low for default effect sizes. As we calculated above, the Type II error probability for $d = 0.20$ with $N = 500$ and $\alpha = 0.05$ is approximately $\beta = 0.278$. This notable imbalance between α and β surely does not reflect the relative seriousness of these types of errors. Thus, when dealing with limited resources, a better strategy may be to define the critical value such that we can minimize both error probabilities. This is the aim of a *compromise power analysis* (Erdfelder, 1984). With a compromise power analysis, we calculate the critical value for the test statistic such that the ratio of the test's error probabilities β/α reflects the relative seriousness of these errors. This may result in non-standard values for both error probabilities; however, there is no rationale for maintaining a conventional level for one error probability if that implies an unreasonably large level for the other (Erdfelder et al., 1996). G*Power can perform compromise power analyses. For a given effect size and sample size, it calculates the critical value that minimizes the error probabilities with the desired ratio. In our example, with $d = 0.20$ and $N = 500$, a compromise power analysis with $\beta/\alpha = 1$ results in the adjusted error probabilities $\alpha = \beta = 0.132$.

The value of a compromise power analysis becomes even more apparent in the (admittedly less common) case that researchers analyze extremely large samples. Relying on conventional α levels may result in an unreasonably large power to detect even extremely small, practically irrelevant effects. Assume that, in our example, we did not sample 500 or 1,084 observations, but 2,000. For $\alpha = 0.05$, our test would have a power of $1 - \beta = 0.998$ to detect an effect of size $d = 0.20$. Again, the implied


Power Plot
Graph
Table
t tests - Means: Difference between two independent means (two groups)
Tail(s) = One. Total sample size = 500. Allocation ratio N2/N1 = 1
Power (1-β err prob)
1
0.8
0.6
0.4
0.2
0
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
Effect size d
α err prob
= 0.1
= 0.075
= 0.05
Draw plot

Parameters

Plot (on y axis) Power (1-β err prob) ☑ with markers □ displaying the values in the plot

as a function of Effect size d from 0.1 in steps of 0.01 through to 0.5

Plot 3 graph(s) interpolating points

with α err prob from 0.05 in steps of 0.025

and Total sample size at 500

Figure 6.2 *Plotting a power curve for different Type I error probabilities in G*Power (sensitivity analysis).*

ratio $\beta/\alpha = 0.04$ would not reflect the seriousness of these errors. Thus, it would be much more reasonable to balance the error rates by means of a compromise analysis in G*Power, resulting in $\alpha = \beta = 0.013$.

So far, we have focused on the traditional notion that statistical procedures require samples of a fixed, predefined size. It is well known that, with classic test procedures, optionally increasing the sample if one observes a non-significant result may inflate the overall Type I error probability (e.g., Anscombe, 1954). Peeking at data during sampling and optionally terminating when a significant result occurs has been identified as a questionable research practice that compromises error probability control, unless one adjusts the critical value a priori to control for interim peeks (Sagarin et al., 2014). However, the statistical toolbox contains a class of procedures that are specifically designed to monitor the data without inflating error rates. This class is known as sequential analysis.

In sequential analysis, we analyze the data during sampling and terminate as soon as a predefined criterion is reached. Importantly, the criterion is defined such that the overall error probabilities of the procedure can be controlled. The benefit of sequential analysis is that the researcher can terminate the sampling process as soon as the data show an informative result, which leads, on average, to a substantial saving in required observations of up to 50 percent (Wetherill, 1975). Thus, sequential analysis constitutes an important alternative to classical analysis particularly in cases where resources are limited, and researchers want to reduce sample sizes without compromising error control.

Although the first sequential procedures were developed almost a century ago (Barnard, 1946; Wald, 1947), they have been mostly ignored in behavioral research. However, a number of sequential test procedures have recently been developed or rediscovered: sequential probability ratio tests (e.g., Schnuerch & Erdfelder, 2020; Schnuerch et al., 2020), sequential Bayes factors (Schönbrodt et al., 2017), group-sequential tests (Lakens et al., 2021), the independent segments procedure (Miller & Ulrich, 2020), and curtailed sampling (Reiber et al., 2020). Due to their high efficiency compared with classic analysis and the increasing number of available software tools, sequential analysis will certainly play a more important role in the future.

Preregistration

The key concept of an empirical science is that we only hold on to a theory if it stands the test of experience. The more critical this test, the more we trust in the theory (Mayo, 2018). Recent reports of failed replication attempts in several published studies have shaken public trust in a number of seemingly established theories and phenomena (Pashler & Wagenmakers, 2012). If so many claims have been published that do not hold under critical scrutiny, we must ask ourselves how critically these claims have been tested in the first place. In this context, *questionable research practices* (QRPs) have gained notable attention.

QRPs describe practices that undermine the integrity of the empirical research process. For a long time, there has been a strong bias toward publishing significant results – results that support claimed effects and phenomena (Bakker et al., 2012). This *publication bias* is not only a problem because it fosters a heavily skewed picture of the existing empirical support for and against a certain claim but it also constitutes a clear incentive for researchers to exploit their degrees of freedom to produce significant results (John et al., 2012; Simmons et al., 2011). Popular examples of QRPs are *HARKing* (hypothesizing after the results are known; Kerr, 1998) and *p-hacking* (e.g., peeking at the data without correction or selectively reporting conditions in which results are consistent with the hypothesis).

QRPs contradict the notion of a critical test. Hence, they are much more likely to lead to false, non-replicable results. Unlike overt scientific fraud, however, QRPs may be perceived as some gray area within the norms of good scientific practice, despite their damaging potential (Simmons et al., 2011), thus rendering them more acceptable and prevalent among researchers (John et al., 2012). Not surprisingly, QRPs have been identified as one factor underlying low replication rates (Schimmack, 2020).

One promising way to reduce QRPs is by means of *preregistration* (Nosek et al., 2018). A preregistration is a time-stamped protocol of all tested hypotheses, experimental design (including sample size), and planned analyses of an experiment, which is created and submitted to a central registry *before* data collection (Wagenmakers et al., 2012). One of these registries is the Open Science Framework (https://osf.io), an open-source repository that can be used to store both preregistrations as well as study material (e.g., stimuli, data, and analysis code; Foster & Deardorff, 2017). The advantage of a preregistered protocol is that it limits researchers' degrees of freedom after data collection, such as HARKing, unplanned optional stopping, selective reporting, or switching from two-tailed to one-tailed tests when this has not been stated explicitly prior to data collection. Moreover, it forces researchers to carefully think about and justify the parameters of their study. Thereby, preregistration may improve the quality and credibility of social and behavioral research (Nosek et al., 2018).

Despite their positive effects, preregistrations cannot eliminate all degrees of freedom, and the effectiveness of preregistering one's research critically depends on the specificity of the preregistered protocol (Bakker et al., 2020). Moreover, preregistrations cannot prevent the problem of publication bias (Scheel et al., 2021). Even though a study has been preregistered, it may still be discarded by a disappointed researcher or rejected by unfavorable reviewers due to unexpected or non-significant results.

A strategy that may provide a better remedy for publication bias is a *registered report* (Chambers & Tzavella, 2020; Greve et al., 2013). A registered report is submitted to a journal for peer review before data collection. At the time of submission, it only contains the theoretical background, the hypotheses, and a detailed description of the planned study design and analyses. The paper may then be accepted based on this study protocol. If the paper is accepted, it will be published irrespective of the results (i.e., as long as the authors adhere to the accepted protocol). This format has many advantages compared with the traditional publication pipeline. Researchers receive reviewer feedback at an earlier stage, enabling them to address concerns and improve the study design. Moreover, it prevents publication bias because the publication decision is made independently of whether the results are significant or not. Thus, incentives for researchers are shifted from producing significant results toward formulating convincing hypotheses and building informative studies. Although the publication format is relatively new, the number of registered reports steadily increases, and the results are promising (Scheel et al., 2021).

Conclusion

Empirical studies are the backbone of social and behavioral research. They may be used to explore new phenomena, develop theories, and critically test hypotheses. Empirical studies can only foster scientific progress; however, if they afford the necessary informativeness with respect to their underlying research question. In this chapter, we have discussed several steps involved in building an informative study. Each of these steps requires careful consideration that may consume a considerable amount of time. This may seem daunting and recourse to simple heuristics (e.g., when formulating a statistical hypothesis or choosing a sample size) may appear attractive. However, like a lumberjack who will have wasted his time trying to cut down a tree with a blunt axe, conducting a poorly built, uninformative study will waste not only the researcher's but also the participants' time as well as valuable public resources. Thus, these resources are better spent on properly building an informative study, and the results will benefit the researcher as well as the entire field of social and behavioral research.

References

Aberson, C. L. (2019). *Applied Power Analysis for the Behavioral Sciences*, 2nd ed. Routledge, Taylor & Francis Group.

Anscombe, F. J. (1954). Fixed-sample-size analysis of sequential observations. *Biometrics*, *10*, 89–100. https://doi.org/10.2307/3001665

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*(6), 423–437. https://doi.org/10.1037/h0020412

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. https://doi.org/10.1177/1745691612459060

Bakker, M., Veldkamp, C. L. S., van Assen, M. A. L. M., et al. (2020). Ensuring the quality and specificity of preregistrations. *PLOS Biology*, *18*(12), e3000937. https://doi.org/10.1371/journal.pbio.3000937

Barnard, G. A. (1946). Sequential tests in industrial statistics. *Supplement to the Journal of the Royal Statistical Society*, *8*(1), 1–26. https://doi.org/10.2307/2983610

Bredenkamp, J. (1972). *Der Signifikanztest in der psychologischen Forschung [The Test of Significance in Psychological Research]*. Akademische Verlagsgesellschaft.

Bredenkamp, J. (1980). *Theorie und Planung psychologischer Experimente [Theory and Planning of Psychological Experiments]*. Steinkopff.

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, *2*(16), 1–38. https://doi.org/10.5334/joc.72

Brysbaert, M. & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1), 9. https://doi.org/10.5334/joc.10

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*(4), 297–312. https://doi.org/10.1037/h0040950

Campbell, J. I. D. & Thompson, V. A. (2012). MorePower 6.0 for ANOVA with relational confidence intervals and Bayesian analysis. *Behavior Research Methods*, *44*(4), 1255–1265. https://doi.org/10.3758/s13428-012-0186-0

Chambers, C. D. & Tzavella, L. (2020). The past, present, and future of registered reports [Preprint]. *MetaArXiv*. https://doi.org/10.31222/osf.io/43298

Champely, S. (2020). *pwr:* Basic functions for power analysis [Manual]. Available at: https://CRAN.R-project.org/package=pwr.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Erlbaum.

Cooper, E. H. & Pantle, A. J. (1967). The total-time hypothesis in verbal learning. *Psychological Bulletin*, *68*(4), 221–234. https://doi.org/10.1037/h0025052

Craik, F. I. M. & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29. https://doi.org/10.1177/0956797613504966

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*(3), 193–242. https://doi.org/10.1037/h0044139

Erdfelder, E. (1984). Zur Bedeutung und Kontrolle des beta-Fehlers bei der inferenzstatistischen Prüfung log-linearer Modelle [On importance and control of beta errors in statistical tests of log-linear models]. *Zeitschrift für Sozialpsychologie*, *15*, 18–32.

Erdfelder, E. (1994). Erzeugung und Verwendung empirischer Daten [Generation and Use of Empirical Data]. In T. Herrmann & W. Tack (eds.), *Methodologische Grundlagen der Psychologie* (Vol. 1, pp. 47–97). Hogrefe.

Erdfelder, E. & Bredenkamp, J. (1994). Hypothesenprüfung [Hypothesis Testing]. In T. Herrmann & W. Tack (eds.), *Methodologische Grundlagen der Psychologie* (Vol. 1, pp. 604–648). Hogrefe.

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, *28*(1), 1–11. https://doi.org/10.3758/BF03203630

Falk, A. & Heckman, J. J. (2009). Lab experiments are a major source of knowledge in the social sciences. *Science*, *326*(5952), 535–538. https://doi.org/10.1126/science.1168244

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Foster, E. D. & Deardorff, A. (2017). Open Science Framework (OSF). *Journal of the Medical Library Association*, *105*(2), 203–206. https://doi.org/10.5195/JMLA.2017.88

Fu, Q., Hoijtink, H., & Moerbeek, M. (2021). Sample-size determination for the Bayesian *t* test and Welch's test using the approximate adjusted fractional Bayes factor. *Behavior Research Methods*, *53*(1), 139–152. https://doi.org/10.3758/s13428-020-01408-1

Gelman, A. & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

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). Erlbaum.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606. https://doi.org/10.1016/j.socec.2004.09.033

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Greve, W., Bröder, A., & Erdfelder, E. (2013). Result-blind peer reviews and editorial decisions: A missing pillar of scientific culture. *European Psychologist*, *18*(4), 286–294. https://doi.org/10.1027/1016-9040/a000144

Guven, C. & Lee, W.-S. (2015). Height, aging and cognitive abilities across Europe. *Economics & Human Biology*, *16*, 16–29. https://doi.org/10.1016/j.ehb.2013.12.005

Hays, W. L. (1963). *Statistics*. Holt, Rinehart and Winston.

Heck, D. W. & Erdfelder, E. (2019). Maximizing the expected information gain of cognitive modeling via design optimization. *Computational Brain & Behavior*, *2*(3–4), 202–209. https://doi.org/10.1007/s42113-019-00035-0

Highhouse, S. & Gillespie, J. Z. (2009). Do samples really matter that much? In C. E. Lance & R. J. Vandenberg (eds.), *Statistical and Methodological Myths and Urban Legends: Doctrine, Verity and Fable in the Organizational and Social Sciences* (pp. 247–265). Routledge, Taylor & Francis Group.

Jager, J., Putnick, D. L., & Bornstein, M. H. (2017). More than just convenient: The scientific merits of homogeneous convenience samples. *Monographs of the Society for Research in Child Development*, *82*(2), 13–30. https://doi.org/10.1111/mono.12296

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573–603. https://doi.org/10.1037/a0029146

Kumle, L., Võ, M. L.-H., & Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods*, *53*, 2528–2543. https://doi.org/10.3758/s13428-021-01546-0

Lakatos, I. (1978). *The Methodology of Scientific Research Programmes*. Cambridge University Press.

Lakens, D. (2021). The practical alternative to the *p* value is the correctly used *p* value. *Perspectives on Psychological Science*, *16*(3), 639–648. https://doi.org/10.1177/1745691620958012

Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, *8*(1), 33267. https://doi.org/10.1525/collabra.33267

Lakens, D. & Caldwell, A. R. (2021). Simulation-based power analysis for factorial analysis of variance designs. *Advances in Methods and Practices in Psychological Science*, *4* (1), 251524592095150. https://doi.org/10.1177/2515245920951503

Lakens, D., Adolfi, F. G., Albers, C. J., et al. (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168–171. https://doi.org/10.1038/s41562-018-0311-x

Lakens, D., Pahlke, F., & Wassmer, G. (2021). Group sequential designs: A tutorial [Preprint]. *PsyArXiv*. https://doi.org/10.31234/osf.io/x4azm

Landers, R. N. & Behrend, T. S. (2015). An inconvenient truth: Arbitrary distinctions between organizational, mechanical turk, and other convenience samples. *Industrial and Organizational Psychology*, *8*(2), 142–164. https://doi.org/10.1017/iop.2015.13

Leatherdale, S. T. (2019). Natural experiment methodology for research: A review of how different methods can support real-world research. *International Journal of Social Research Methodology*, *22*(1), 19–35. https://doi.org/10.1080/13645579.2018.1488449

Lin, H., Werner, K. M., & Inzlicht, M. (2021). Promises and perils of experimentation: The mutual-internal-validity problem. *Perspectives on Psychological Science*, *16*(4), 854–863. https://doi.org/10.1177/1745691620974773

Mayo, D. G. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge University Press.

Meiser, T. (2011). Much pain, little gain? Paradigm-specific models and methods in experimental psychology. *Perspectives on Psychological Science*, *6*(2), 183–191. https://doi.org/10.1177/1745691611400241

Miller, J. & Ulrich, R. (2020). A simple, general, and efficient method for sequential hypothesis testing: The independent segments procedure. *Psychological Methods*, *26*(4), 486–497. https://doi.org/10.1037/met0000350

Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E.-J. (2014). Why hypothesis tests are essential for psychological science: A comment on Cumming (2014). *Psychological Science*, *25*, 1289–1290. https://doi.org/10.1177/0956797614525969

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Pashler, H. & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science*, *9*, 319–332. https://doi.org/10.1177/1745691614528519

Perugini, M., Gallucci, M., & Costantini, G. (2018). A practical primer to power analysis for simple experimental designs. *International Review of Social Psychology*, *31*(1). https://doi.org/10.5334/irsp.181

Popper, K. R. (1968). *The Logic of Scientific Discovery*, 3rd ed. Hutchinson.

Reiber, F., Schnuerch, M., & Ulrich, R. (2020). Improving the efficiency of surveys with randomized response models: A sequential approach based on curtailed sampling. *Psychological Methods*, *27*(2), 198–211. https://doi.org/10.1037/met0000353

Reichenbach, H. (1938). *Experience and Prediction: An Analysis of the Foundations and the Structure of Knowledge*. University of Chicago Press. https://doi.org/10.1037/11656-000

Roberts, S. & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358–367. https://doi.org/10.1037/0033-295X.107.2.358

Rouder, J. N., Morey, R. D., & Wagenmakers, E.-J. (2016). The interplay between subjectivity, statistical practice, and psychological science. *Collabra*, *2*, 1–12. https://doi.org/10.1525/collabra.28

Rouder, J. N., Schnuerch, M., Haaf, J. M., & Morey, R. D. (2022). Principles of model specification in ANOVA designs. *Computational Brain & Behavior*. https://doi.org/10.1007/s42113-022-00132-7

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237. https://doi.org/10.3758/PBR.16.2.225

Sackett, P. R. & Larson Jr., J. R. (1990). Research strategies and tactics in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (eds.), *Handbook of Industrial and Organizational Psychology, Volume 1*, 2nd ed. (pp. 419–489). Consulting Psychologists Press.

Sagarin, B. J., Ambler, J. K., & Lee, E. M. (2014). An ethical approach to peeking at data. *Perspectives on Psychological Science*, *9*(3), 293–304. https://doi.org/10.1177/1745691614528214

Scheel, A. M., Schijen, M. R. M. J., & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with registered reports. *Advances in*

Methods and Practices in Psychological Science, *4*(2), 251524592110074. https://doi.org/10.1177/25152459211007467

Schimmack, U. (2020). A meta-psychological perspective on the decade of replication failures in social psychology. *Canadian Psychology/Psychologie Canadienne*, *61*(4), 364–376. https://doi.org/10.1037/cap0000246

Schnuerch, M. & Erdfelder, E. (2020). Controlling decision errors with minimal costs: The sequential probability ratio *t* test. *Psychological Methods*, *25*(2), 206–226. https://doi.org/10.1037/met0000234

Schnuerch, M., Erdfelder, E., & Heck, D. W. (2020). Sequential hypothesis tests for multinomial processing tree models. *Journal of Mathematical Psychology*, *95*, 102326. https://doi.org/10.1016/j.jmp.2020.102326

Schönbrodt, F. D. & Stefan, A. M. (2019). BFDA: An R package for Bayes factor design analysis (version 0.5.0) [Manual]. Available at: https://github.com/nicebread/BFDA.

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2), 322–339. https://doi.org/10.1037/met0000061

Schram, A. (2005). Artificiality: The tension between internal and external validity in economic experiments. *Journal of Economic Methodology*, *12*(2), 225–237. https://doi.org/10.1080/13501780500086081

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Ulrich, R., Miller, J., & Erdfelder, E. (2018). Effect size estimation from *t*-statistics in the presence of publication bias: A brief review of existing approaches with some extensions. *Zeitschrift für Psychologie*, *226*, 56–80. https://doi.org/10.1027/2151-2604/a000319

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*(6), 491–498. https://doi.org/10.1016/j.jmp.2010.07.003

Vanpaemel, W. & Lee, M. D. (2012). Using priors to formalize theory: Optimal attention and the generalized context model. *Psychonomic Bulletin & Review*, *19*(6), 1047–1056. https://doi.org/10.3758/s13423-012-0300-4

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*, 779–804. https://doi.org/10.3758/BF03194105

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632–638. https://doi.org/10.1177/1745691612463078

Wald, A. (1947). *Sequential Analysis*. Wiley.

Wetherill, G. B. (1975). *Sequential Methods in Statistics*, 2nd ed. Chapman and Hall.

7 Analyzing Data

Roger Watt and Elizabeth Collins

Abstract

This chapter provides an overview of the processes that are commonly used for analyzing data. Our intention is to explain what these processes achieve and why they are done. Analyzing data goes through four stages. For each stage, we explain the most important concept and then explain the practical steps that are involved. This begins with the data themselves as variables. Next, we move on to describing the data, their variance and covariance, with linear models. Next, we cover interpreting effects and focus on effect sizes. We end with a discussion of inferences about the population and how the presence of uncertainty has to be taken into account in reaching conclusions.

Keywords: Variables, Variance, Covariance, Sampling, Linear Models, Effect Sizes, Uncertainty, Causality

Introduction

Wherever one looks, there is variability in how people are and how they respond to different situations. Data captures that variability, and their statistical analysis brings them to life.

Statistics is often seen as a mystery. The beginner is faced with a multiplicity of tests and things to take into account. The general impression is that the a priori odds of doing an analysis correctly are small. Nearly all this problem is simply clutter that has accumulated over time and not yet been cleaned out. For example, the *t*-test was a very early statistical test. It could have been replaced by the one-way analysis of variance (ANOVA) test when that was introduced, but the two tests were kept despite the fact that both give exactly the same result. Both these tests could have been replaced by the General Linear Model when that was introduced but again all three continued to exist side by side. In this chapter, we remove the mystery by explaining how to understand rather than instructing how to do.

When we speak of data, what do we mean? In the social and behavioral sciences, data are typically a set of observations made on a sample taken from a population. The sample represents the whole population with the hope that it will contain the patterns of variability that are to be found in the population.

Data analysis examines variability in data, exploring what lies behind. Data are organized into *variables* of interest, with one value for each member of the sample. The variables are ways in which members of the population or the situations they are in differ. For example, there is variability in how people comply with instructions. We may have the thought that differing degrees of social pressure may be involved in this variability in compliance. Our data will need variables for compliance and for social pressure to capture each person's own characteristic value for compliance and their experience of social pressure. We can rely on observing naturally occurring variations in social pressure or we can do a social pressure experiment (assuming we find a way of doing it ethically) by introducing a varying situation – participants are either in low or high social pressure.

Until they have been analyzed, the data themselves are valueless. At the most, they record that, at one specific moment in a specific situation, a specific person responded in a specific way. They are a collection of anecdotes. Data are analyzed to reach provisional answers to questions of theoretical or practical importance about a population.

Using data from a sample to answer a question about a population can only lead to an uncertain inference. Data analysis must also estimate the degree of uncertainty involved in that inference. It is extremely helpful, and not a little magical, that the same sample of data can fulfill both objectives – an inference about the population and an estimate of the degree of uncertainty in that inference. The same data can provide an answer to the question "What effect is there?" and, simultaneously, to the question "How reliable is this answer?".

There may be patterns of association between the variables in data. Some of these associations will be interesting, others not; some will be accidental, others not. These associations can be expressed as shared variability. The question "Are people more compliant with instructions when they are under high social pressure?" asks whether the variability that people show in their degree of compliance with instructions is associated with the variability in social pressure that they experience. The two scatter plots in Figure 7.1 show data that have the same overall variability. However, in Figure 7.1b the variability in social pressure and compliance is shared and we might conclude that there is an association between social pressure and compliance. Finding patterns like this is the goal of data analysis.

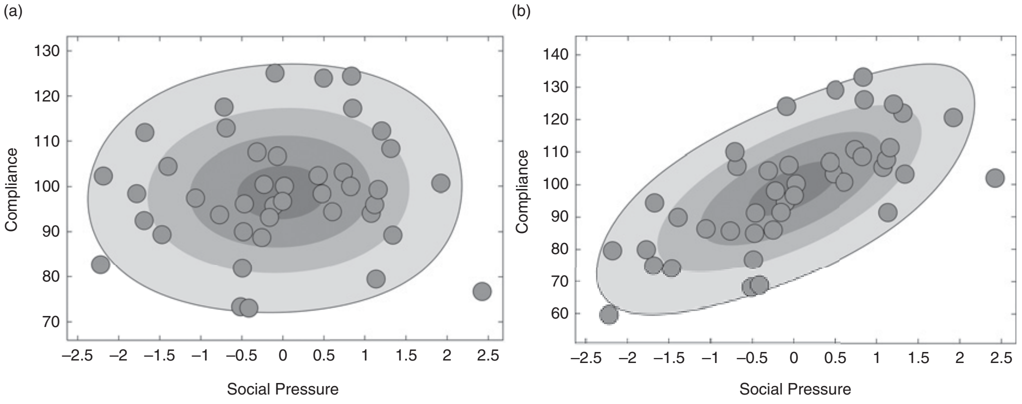


Figure 7.1 *Two scatter plots with no effect (a) and a substantial effect (b). The variability of each variable taken on its own (i.e., the variance) is the same in both (a) and (b). In (b), but not (a), the variables share some of that variability. This is called covariance – they vary together.*

There are four questions to address in the analysis of data.

(1) **How good are the data?** We must ensure that the quality of the data is as good as possible and that any aspects of the data that are less than ideal are identified, often by visual inspection of the data, and, if possible, rectified. *Do the data appear to adequately capture social pressure and compliance?*

(2) **What patterns are present in the sample?** Then we describe the patterns of variability in the data itself. This will identify associations between variables in the data. This description of the sample is numerically exact – it is what the specific sample is. *Do the data show an association between social pressure and compliance?*

(3) **How do we interpret the effects of different variables?** We convert the effects that we see into standardized forms so that we can easily understand the relative importance of different effects. *What is the effect size for social pressure on compliance?*

(4) **How reliable is the generalization to the population?** The fourth step is to use the description of the sample to make an uncertain inference about the population it was drawn from. This inference is an extrapolation (i.e., a step beyond the facts), and we must evaluate its uncertainty. *How much can we say about compliance with instructions in the wider population, given the limitations of our sample?*

Analyzing Data 1: The Sample

We have a sample. What variability does it contain?

Concept 1: Variables, Variance, and Covariance

The simplest form of variability produces different categories (e.g., eye color: blue, brown, green; degree title: BA, BSc). These categories may also be different situations, such as time point (before, after) or experimental group (treatment, control).

A *categorical* (sometimes called *nominal*) variable is when the only comparison between two values is whether they are the same or different. The values are different categories, also known as cases, groups, or levels. In our example, ‘social pressure group’ would be a categorical variable if we created an experiment where participants were assigned to low or high social pressure groups.

Sometimes it is possible to put the variability into a meaningful order. This cannot be done for eye color, but if we categorize people as being in the top third, middle third, or bottom third for exam grades, then there is a meaningful order.

An *ordinal* variable has an ordering that allows more/less comparisons between values as well as same/different. With an ordinal variable, it is safe to say that one value is higher than another but not necessarily safe to say by how much it is higher. A ranking scale from 0 (“not at all”) to 7 (“always”) is ordinal. The example we have just given, of people being in the top, middle, or bottom third of exam grades, is an example of an ordinal variable that resembles categories.

Finally, numbers can be used for many variables where the numbers would relate to some measurable quantity (e.g., a score on a personality test or an actual exam grade), sometimes referred to as a *scale*.

An *interval* variable has the property that addition and subtraction are meaningful. With this property, we can say how much higher or lower one value is compared to another. The standard IQ scale is an example of an interval value: a difference of 10 points means the same anywhere in the scale.

A *ratio* variable has the additional property that multiplication is also meaningful, including multiplication by zero so that a zero value equates to an absence of the characteristic being measured. A bank balance is a ratio variable – zero means an absence of funds and negative values carry a very different meaning from positive values.

It is useful to be able to calculate the amount of variability in the values of a variable. The most common measure when the variable is a quantity (interval/ ratio variable) is called the *variance*. Variance is a measure of how much a set of values are different from each other. When the variable is categorical, the equivalent measure of variability is *deviance*. Deviance is a measure of how many values are different from each other. If every value in the sample is the same, then the variance or deviance is zero. The more different from each other the participants are, the higher the variance or deviance.

If we take any two people, *i* and *j*, with two different values, x_i and x_j for some variable, we can calculate the arithmetic difference between those two values. If we have *n* people, then we have $n \times n$ possible pairs of people (this includes a person and themself). So, we have $n \times n$ differences between pairs of people. We square these differences (this removes the sign):

$$(x_i - x_j)^2$$

and find the mean of all the possible squared differences:

$$\frac{1}{n}\sum_i \frac{1}{n}\sum_j (x_i - x_j)^2$$

Variance is half of this because each difference gets counted twice: $\left(x_i - x_j\right)$ and $\left(x_j - x_i\right)$:

$$var(x) = \frac{1}{2}\frac{1}{n}\sum_i \frac{1}{n}\sum_j \left(x_i - x_j\right)^2$$

Variance can also be calculated by

(i) taking the deviation (difference) of each data point from the mean value for the variable
(ii) squaring each deviation
(iii) summing to produce the sum of squared deviations

(iv) dividing by n to get the mean squared difference:

$$var(x) = \frac{1}{n}\sum_{i}(x_i - \overline{x})^2$$

The sum of squared deviations, often abbreviated to SumSq or SSq, will re-appear later in this chapter.

How about associations between two variables with values x_i and y_i? Returning to Figure 7.1, the variance of each individual variable is the same in both scatter plots. In Figure 7.1b, the variables share some variability – higher social pressure (x) is typically associated with higher compliance (y). There is a quantity, *covariance*, that captures this. It is similar in many respects to variance. The two formulae are given here:

$$var(x) = \frac{1}{n}\sum_{i}(x_i - \overline{x})(x_i - \overline{x})$$

$$cov(x) = \frac{1}{n}\sum_{i}(x_i - \overline{x})(y_i - \overline{y})$$

If the two variables, x and y, have no association (are unrelated), that would result in their covariance being zero. If there was an association between the two variables they would have a non-zero covariance. Therefore, a simple form of explaining data analysis is to say that it looks for the pattern of covariance between the variables.

Variance has two very useful properties. When we add together the values from two independent variables, the result is a new variable that has a variance given by the sum of the two variances of the variables we are adding. If we make a new variable by adding a constant to each value of an existing variable, then its variance is not changed. If we change a variable by multiplying each value by a constant, k, then the variance is itself multiplied by k^2:

Rule 1: independent variances add: $var(x + y) = var(x) + var(y)$
Rule 2: variances scale by squares: $var(k \times x) = k^2 var(x)$

These will allow us to say, for example, that the variance in compliance is 225. The values themselves in Figure 7.1 have a range between 70 and 130. So they are spread over a range of about 60 from the smallest to the largest and they have a mean of about 100. A typical deviation is around 15 (some are near enough 0 and some deviations are as much as 30). So a mean (i.e., typical) squared deviation is (15 squared) = 225. The portion that appears to be caused by social pressure is 45, leaving 180 unaccounted for and indicating that variance in social pressure accounts for 20% of the variance in compliance. Deviance has a similar property and is used in a similar way.

There is a more general form of Rule 1 when x and y are related:

Rule 3: $var(x + y) = var(x) + var(y) + 2cov(x, y)$

Practical Step 1: Inspect the Sample for Errors

The first practical step is to inspect the data for any indication that the sampling was poor. Ideally, the sampling was completely random; we have to assume this to estimate uncertainty. In practice, some compromises will have been made in the sampling method, and the data may contain errors. Visual inspection of the data is an important step as it might reveal any defects that have resulted.

Key to this is to understand that any data analysis is very sensitive to data points at the extremes. Figure 7.2 shows a sample that has a positive relationship between the two variables. The data points marked in white, relatively extreme in both variables, are the ones that contribute most to the analysis of that relationship. These are said to have the highest *leverage*. If the sampling method had not been able to capture points with these extreme values, then the relationship would appear weaker and would not be a fair estimate of the population effect.

The most important potential problems in data are these:

- incompleteness: not the whole population was available to be sampled
- not a single population: points outside the population were sampled
- non-independence: the choice of some data points depended on other data points

Finding out whether the sample has a limited range, as in Figure 7.3, is not simple. Sometimes the population distribution of a variable is already well known. For example, the standard measure of intelligence, IQ, has a known population. If a sample of IQ values are noticeably different from what is expected, then the

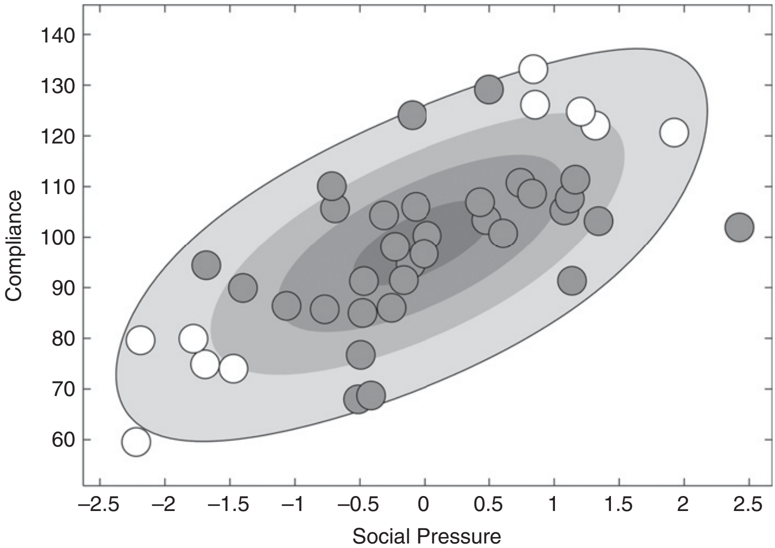


Figure 7.2 *Simulated data showing a hypothetical relationship between social pressure and compliance. The relationship in the sample is small but not zero. Detecting it depends on the data points marked in white; without these, the relationship would be negligible.*

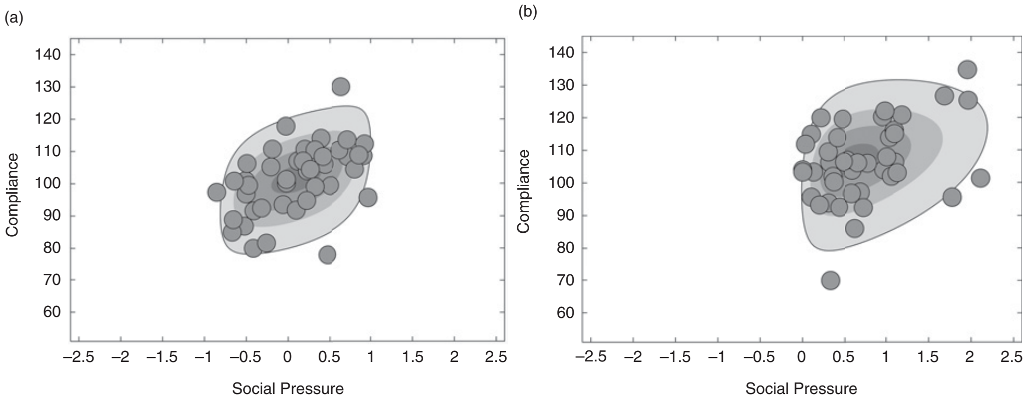


Figure 7.3 *Some data with incomplete sampling: (a) sampling that missed both tails of the social pressure distribution and (b) sampling that has missed the left half of the distribution. Comparing these with Figure 7.2 shows how important completeness is.*

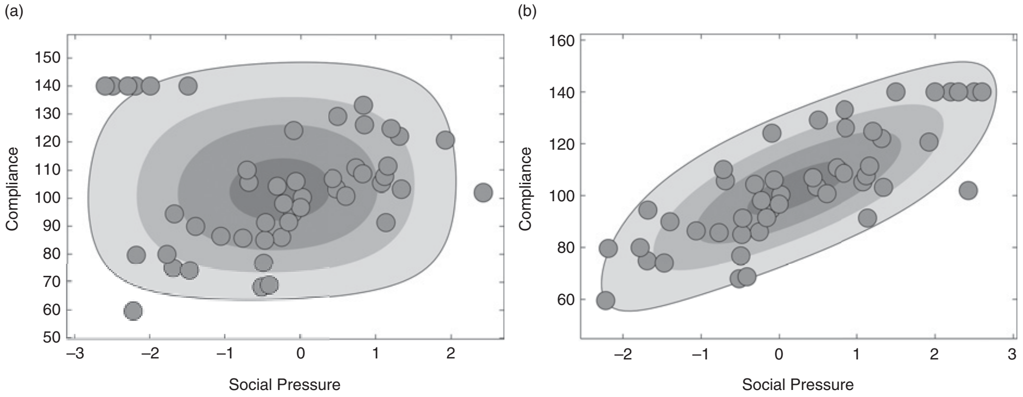


Figure 7.4 *Two scatter plots of data: (a) the original data with some outliers (top left) that eradicate the effect; (b) the outliers (top right) amplify the effect.*

sampling range may have been limited. More generally, one can often assume that the distribution of values should be close to the normal distribution. If a sample is drawn too heavily from the middle of the distribution, then its distribution will show up too flat (lower than expected kurtosis). If one or other tail of the distribution is under-sampled, then the sample distribution will be stretched more in one direction than the other (a stronger skew than expected).

Extraneous data points that do not belong to the population in question can affect data analysis. These may show up as outliers – points with implausible values, although extraneous data points are not limited to extreme values. However, extreme values are both (i) easier to detect and (ii) potentially more problematic. Figure 7.4 shows a set of data with outliers added to either the left or the right with quite

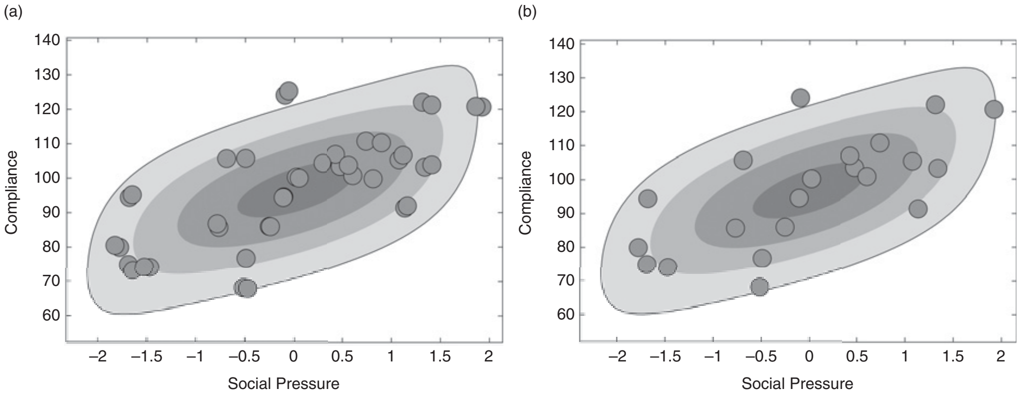


Figure 7.5 *The sample of data shown here have high non-independence. This is visible in this case as small, localized clusters of data points. In (b) the data have just one from each small cluster (so half as many data points) and, as can be seen, this is really what the data in (a) contain. The extra non-independent data do not add anything except the illusion of higher precision.*

different consequences. There is considerable guidance for the treatment of outliers, but it is all necessarily ad hoc. Removing outliers is especially problematic if the only indication that a data point does not belong is the very subjective observation that it does not appear to fit. Our advice is only to remove a data point if a clear error can be identified.

If the members of a sample are not independent of each other, the uncertainty will be underestimated, giving more confidence in the result than is justified. Non-independence can appear as clustering of data points, as in Figure 7.5, but is difficult to detect and is also best avoided to start with.

Analyzing Data 2: Describing the Data

Now that we have examined the data, we move on to describe how the variables relate to one another.

Concept 2: Linear Models

We will use a *model* to describe any meaningful patterns in the sample. A model is a description of the data that aims to capture its salient features. A common model in social and behavioral sciences is a *linear model*, called that because the different parts are *added* together.

A simple example of a model is the statement that the observed values of compliance in a sample are the sum of two different sources: a contribution from social pressure and the net contribution of all the remaining unknown factors that, when lumped together, are called the *residual*. We start with a simple equation:

$$\text{Equation 7.1: } compliance_i = \ a + b \times socialPressure_i + residual_i$$

Equation 1 says that the compliance for person *i* is made up of a fixed quantity, *a*, plus an amount that is their own *socialPressure* multiplied by another fixed quantity, *b*, plus another (unknown) residual amount specific to that person. The fixed quantities, called *coefficients*, *a* and *b*, are the same for everybody. The variables, on the other hand, *compliance*, *socialPressure,* and *residual* have different values for each participant *i*. The residual contains all the other influences on compliance that haven't been investigated.

To write down the model itself, the *i* subscripts can be left out since it now applies to everyone. Since the model is an idealization of the data, the residual term is not written down either. Without the residual, we use a ← sign instead of the = sign:

Model 1: $compliance \leftarrow a + b \times socialPressure$

It is conventional to call the outcome variable on the left (i.e., *compliance*) the *response variable* and the variables on the right the *predictor variables*. The coefficient *b* sets how much the observed compliance is affected by the observed social pressure. If *b* is zero, then social pressure has no effect on compliance; if *b* is greater than zero, more social pressure means more compliance; if *b* is negative, more social pressure means less compliance.

What if a predictor variable has categories not quantities? A simple extension allows the use of categorical predictor variables to make a *general linear model*. For example, an experimental variable to represent the different phases of an intervention, which has the values of before and after, can be included in a model like this:

$$compliance \leftarrow a + b \times (phase = after)$$

What if the response variable is categorical? Another simple extension allows the response variable also to be categorical, and we have a *generalized linear model*. The approach is the same except that a hidden continuous variable is placed between the response variable and the predictors:

$$hidden \leftarrow a + b \times socialPressure$$

$$compliance(yes/no) \leftarrow binary(hidden)$$

The first part of this is a linear model as before; the second part has a probabilistic function *binary ()* that converts the continuous *hidden* to the categorical outcome.

A model can often involve many different terms. For example, here is a model with several predictors:

$$compliance \leftarrow a + b_1 \times socialPressure + b_2 \times extroversion + b_3 \times age$$

The coefficients determine how much each predictor contributes to the response. In theory, we can give the coefficients any values we like, and the residual term can be adjusted to make Equation 7.1 valid. For example, if they were all zero, then the residual values would be just the same as the response values (compliance).

Practical Step 2: Fitting Linear Models

Suitable values for the coefficients are found by a fitting process designed to produce the *best-fitting model*. The best-fitting model is often the model that minimizes the discrepancies between the data and the model predictions (i.e., minimizes the residuals). Specifically, a *least-squares* model is one that finds coefficients that create the smallest sum of squared residuals. That is the model that has the smallest variance of its residuals.

Return to this model:

$$compliance \leftarrow a + b \times social\,Pressure$$

Figure 7.6 shows the sample data from Figure 7.1 with the various components of the model. The gray dots are the actual data points. The least-squares model is shown as the black diagonal line running through the data indicating that compliance goes up when social pressure goes up (in this case). The white squares show the values predicted by the model for each data point; the difference between the two is the residual, shown as the thin vertical lines. If the model line were either less steep or more steep, then a few residuals would decrease in size, but most would increase in size leading to a poorer overall fit.

We can think of different people's compliance value as being the value predicted by the model for them plus a residual. From this, we can see that variability in the model values accounts for some of the variability of the response (compliance) values but not all. We say that the model explains some of the observed variance in the response. This provides a very good way of stating how good a model is – how much of the variance of the response variable is explained by the predictor variables. In this context, the variance of the model is often called the variance explained. In this way, the model gives us three interlinked variances. In the data shown in Figure 7.6, they are:

- var(response): variance of compliance (gray circles) = 225
- var(model): variance of $a + b \times socialPressure$(white squares) $= 58.1$
- var(residuals): variance of residuals (black lines)= 166.9

The variance of the residuals is sometimes called the error variance. The model and the residuals are independent of each other and so their variances add:

$$\text{Equation 7.2}: var(response) = var(model) + var(residuals)$$

which becomes this for our example:

$$var(compliance) = var(a + b\times\ socialPressure) + var(residuals)$$

and for our data, we then have:

$$313.3 = 146.4 + 166.9$$

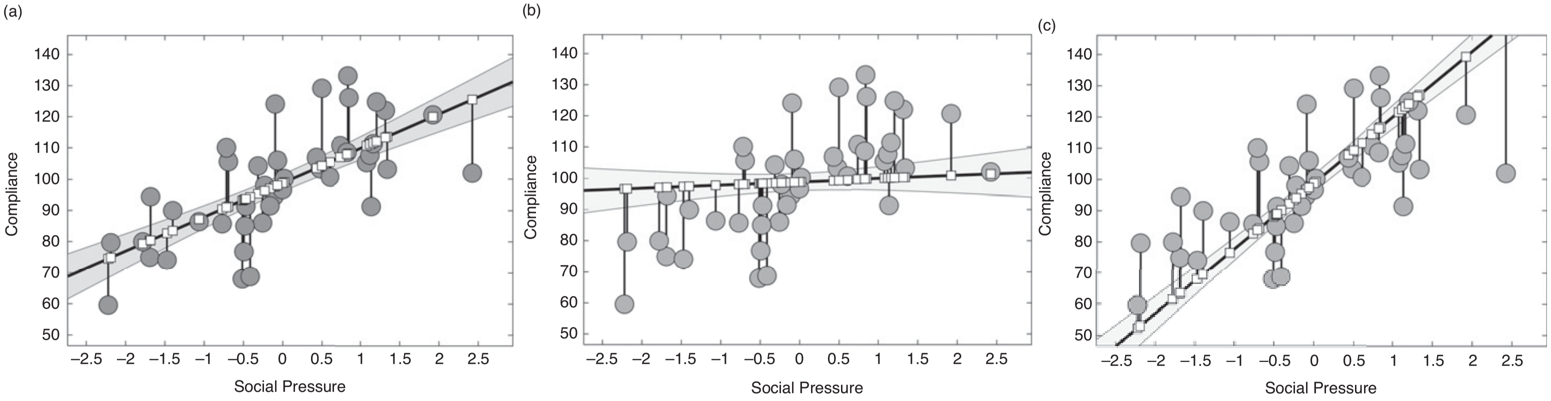


Figure 7.6 *(a) An example of a set of data (circles) with the model values shown (white squares). The residuals are shown in the diagram as the vertical lines between the model value and the data value. Two alternative model lines are either too shallow (b) or too steep (c).*

which says that the variance of compliance is 313, made up of 146 from social pressure and 167 left over as residual.

In this case, with just one predictor variable, there is a simple formula to calculate the coefficients, *a* and *b*:

$$a = mean(compliance)$$

$$b = \frac{cov(compliance, socialPressure)}{var(socialPressure)}$$

The first coefficient, *a*, is rarely of interest as it doesn't say anything about the relationship between the predictors and the response variable. The second coefficient, *b*, is important as it does specify how strong the relationship is. It is very convenient that it has a simple formula involving just variance and covariance.

When the predictor variable is a dichotomous categorical variable, the two coefficients, *a* and *b*, have a simple direct relation to the data. In this case, the two group means are *a* and *a* + *b*.

Something Extra: Interactions

The effects we have discussed so far are all main effects: where one variable affects another – social pressure affects compliance. There is another type of effect called an *interaction*. An interaction effect is where one variable affects the effect of one variable on another. For example, the effect of social pressure on compliance might depend on age – there is no effect of social pressure in young people but a strong effect in older people:

- effect: variable affects a variable
- interaction: variable affects an effect

This interaction is shown in Figure 7.7. To make this graph, the data are split into two age ranges, and each age range has its own line. The shallower slope for the younger age group indicates less effect of social pressure on compliance for this age group.

An interaction is like a switch – the effect of a predictor on the response variable is changed by the value of another predictor. That effect can be switched on or off, from negative to positive, or anything in between. These can easily be accommodated within linear models. Mathematically, they appear as the product of the two (or more) variables involved. This product becomes, mathematically, a new variable made up from the two:

$$compliance \leftarrow a + b \times (socialPressure \times age) + ...$$

This can be slightly re-ordered as:

$$compliance \leftarrow a + (b \times age) \times socialPressure + ...unknown$$

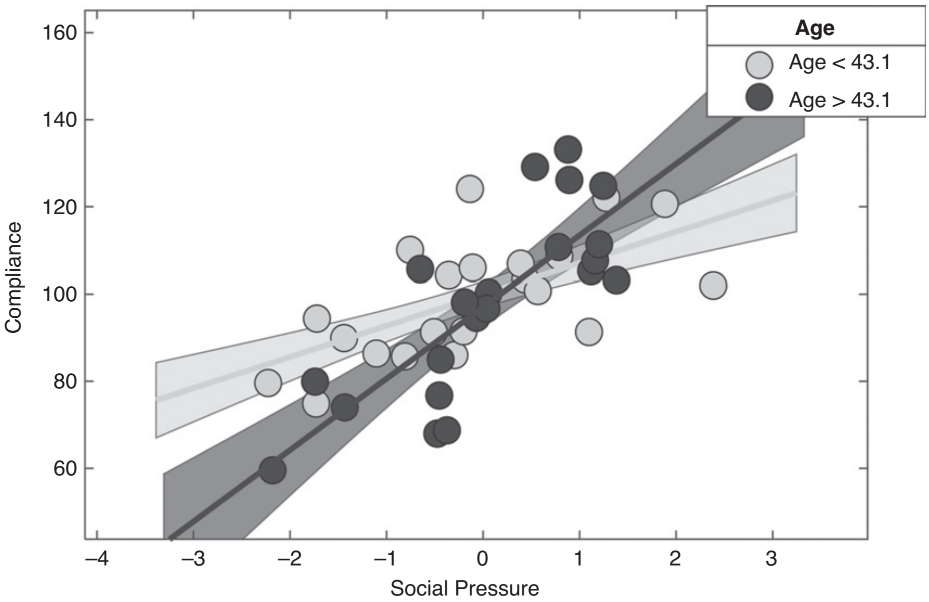


Figure 7.7 *When we add in age as a variable, we see that the effect of social pressure on compliance is higher in older people (dark dots) than in younger people (light dots). This indicates an interaction between social pressure and age.*

By writing it like this, we can see that the effect size that links *socialPressure* to *compliance*, that was previously just *b*, is now $(b \times age)$. The effect size for *socialPressure* is not now a constant; it varies with *age*. Note that the interaction is symmetric – we can just as easily talk of the effect *socialPressure* on the relationship between *age* and *compliance*.

Analyzing Data 3: Interpreting Effects

When we calculate the best-fitting model for the data in Figure 7.6 we find that it has these coefficients:

$$compliance \leftarrow 98.9 + 9.8 \times socialPressure$$

We can just about to see these on the graph. When the social pressure is zero, the compliance in the model will be 98.9 – and it looks like about 100 on the graph. The second coefficient, 9.8, tells us how much additional compliance there will be for an increase in social pressure of 1. So, if we compare compliance for social pressure of 0 and 1, we can see that the difference in compliance is about 10.

What do the coefficients mean? Before we answer that question, briefly consider another example:

$$examGrade \leftarrow 20 + 6 \times hoursStudy$$

which says one hour of study increases the exam grade by 6 points. Here, we can see immediately the practical value of study before an exam; we know what one hour is and we know what 6 grade points are. In our compliance example, an increase of social pressure by 1 increases the compliance by 9.8. Unfortunately, this doesn't tell us whether social pressure is important or not because we don't know what these numbers mean. So, the compliance model is unhelpful. How can we use this model to say something helpful?

Concept 3: Effect sizes

Equation 7.2 shows that the variance of the response variable (compliance) splits into two parts: the part from the model (social pressure) and the part from the residual. These parts are shown in the Venn diagram in Figure 7.8; from this we can see visually the contribution of social pressure to compliance.

This sets the stage for observing two very useful ways of quantifying the association strength between *socialPressure* and *compliance*:

- *Normalized effect size*. We can compare the variance in *compliance*, that is due to *socialPressure*, with the total variance of *compliance*. This is A/(A+B) in the diagram. A normalized effect size is the square root of this comparison. The normalized effect size, *r*, can be calculated directly from the data and the model with this formula:

$$r^2 = \frac{b^2 \times var(socialPressure)}{var(compliance)}$$

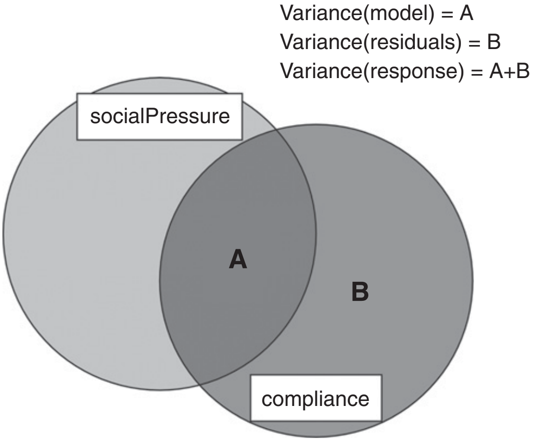


Figure 7.8 *An effect between two variables. The variance of each variable is represented by a circle. Where the circles overlap, there is shared variance.*

or without b:

$$r^2 = \frac{cov(socialPressure, compliance)^2}{var(socialPressure) \times var(compliance)}$$

It can range from 0 to 1 and can be positive or negative.

- *Standardized effect size*: we can compare the variance due to *socialPressure* to the variance of the residuals. This is A/B in the diagram. A standardized effect size is the square root of this comparison. The standardized effect size, f, can be calculated directly from the model with this formula:

$$f^2 = \frac{b^2 \times var(socialPressure)}{var(residuals)}$$

- It can range from 0 to infinity (when the variance of the residuals is zero).

The two are easily interconverted:

$$f^2 = \frac{r^2}{1 - r^2}$$

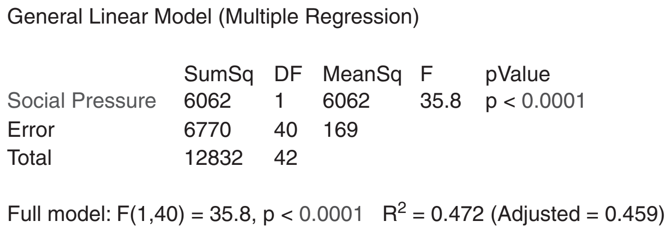
Since both these effect sizes are ratios of parts of the variance of the same response variable, *compliance*, they no longer have arbitrary units. In both cases, if we are told that an effect size is 0.3, we can decide whether this is of practical importance regardless of the numbers used to measure it.

The choice between these two versions of effect size is arbitrary, although a scale from 0 to 1 is easier to manage than one that goes to infinity. There are many different effect sizes in use. Two deserve some attention:

- Correlation coefficient, r: use is typically when there are just two interval variables. In that case, it is exactly equivalent to the normalized effect size as we have derived it.
- Cohen's d: use is typically when the predictor is a dichotomous categorical variable (frequently in an experimental design), and the response is an interval variable. Cohen's d is defined as the difference between the two group means divided by the pooled standard deviation within the groups. Cohen's d is twice the standardized effect size: $d = 2f$.

Practical Step 3: ANOVA

We have seen in the previous sections how a simple linear model can be analyzed to calculate meaningful effect sizes based on splitting the variance of the response to the various different predictor sources. This is the work of an analysis called *ANOVA* (as noted above, this is an acronym for analysis of variance). The procedure of an ANOVA is to take the variance in the response variable and see how it can be



```
General Linear Model (Multiple Regression)

                 SumSq   DF   MeanSq   F      pValue
Social Pressure  6062    1    6062     35.8   p < 0.0001
Error            6770    40   169
Total            12832   42

Full model: F(1,40) = 35.8, p < 0.0001   R2 = 0.472 (Adjusted = 0.459)
```

Figure 7.9 *An ANOVA table for the basic model of compliance and social pressure. Currently, only the SumSq column is of interest. The remaining columns relate to null hypothesis testing, which we consider later.*

partitioned into different parts that can be attributed to the different predictor variables. If we apply an ANOVA to the data, then it produces a standard form of table that has the quantities that we need (shown in Figure 7.9). For now, we are only interested in the column marked SumSq.

The rows of the table in Figure 7.9 correspond to the different terms in the model; error stands for the residuals. The table doesn't give variances, but it gives the sum of squares (SumSq; i.e., the variance times *n* – the number of data points). Notice that the total SumSq is the sum of the individual SumSq for social pressure and error. The rows in the ANOVA table are related to the areas in the Venn diagram in Figure 7.8:

Social pressure	A
Error	B
Total	$A + B$

With these values for SumSq, we can calculate effect sizes. The value of *r* for our sample is the square root of *A* divided by $(A + B) = \sqrt{6062/12832} = 0.69$ and *f* is $\sqrt{6062/6770} = 0.95$.

When there is more than one predictor variable, the analysis has the potential to become more complex because the predictors may themselves be associated. This rarely happens in experimental studies where the two predictors might be two different interventions, carefully managed so that they are fully independent. However, when the predictors are observed variables, then it is quite plausible that there will exist a relationship between them. For example, this model has two predictors: social pressure and extroversion.

Model 2: $compliance \leftarrow a + b_1 \times socialPressure + b_2 \times extroversion$

It is quite plausible that extroverts experience more social pressure as they are more social. So, the two predictors are themselves related.

This more complex situation with Model 2 is shown as a Venn diagram in Figure 7.10. The two predictor variables each overlap the response variable (i.e., they contribute to the variance of the response variable), but they also overlap each other – this is the complication. The variance that *socialPressure* on its own contributes is $A1{+}A12$; *extroversion* on its own contributes is $A2 + A12$. If we are

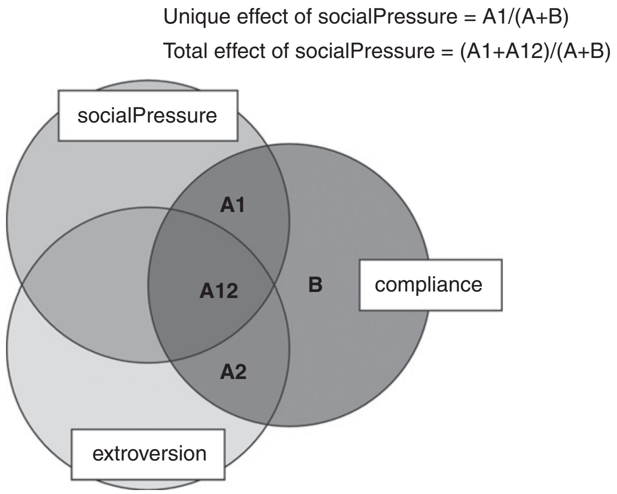

Unique effect of socialPressure = A1/(A+B)
Total effect of socialPressure = (A1+A12)/(A+B)
socialPressure
A1
A12
B
compliance
A2
extroversion

Figure 7.10 *The variance of a response variable and the variance accounted for by two predictors. Since the two predictors are partially correlated, there is an overlap between all three variables.*

not careful, $A12$ gets counted twice, and we think we have accounted for more of the variance in the response than we actually have. To overcome this, the ANOVA introduces two types of effect size:

- The *total effect* of a predictor on a response is the effect seen when only those two variables are analyzed. The total effect of *socialPressure* is $(A1 + A12)/(A + B)$.
- The *unique effect* of a predictor on a response is the effect that only that variable (among the ones in consideration) has on the response. The unique effect of *socialPressure* is $A1/(A + B)$.

These two types of effect will be different from each other unless the predictors are independent of each other.

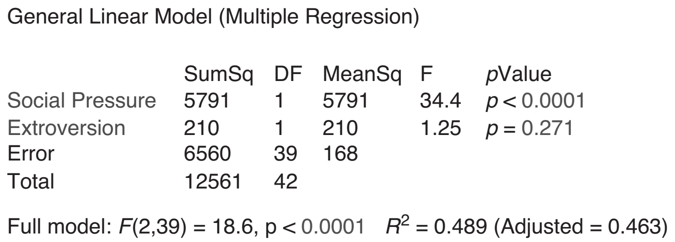
The total effect of a predictor is easy to find; a model is created that has only that predictor and the response in it. The unique effect is a little different. First, a model is created without the particular predictor but with all the other predictors. Then, a second model is created that adds in the predictor of interest. The additional amount of variance in the response variable that is explained by adding in this predictor is then its unique effect. So, to find the unique effect of *socialPressure*, we start with this model:

$$compliance \leftarrow a + c \times extroversion$$

and see how much is gained by switching to this model:

$$compliance \leftarrow a + b \times socialPressure + c \times extroversion$$

The unique effect is of some practical interest. It is often used as a way of *controlling for a covariate*. We observed that social pressure affects compliance, but we want to rule out the possibility that the effect is due to extroversion. So, we



```
General Linear Model (Multiple Regression)

                  SumSq   DF   MeanSq   F      pValue
Social Pressure   5791    1    5791     34.4   p < 0.0001
Extroversion      210     1    210      1.25   p = 0.271
Error             6560    39   168
Total             12561   42

Full model: F(2,39) = 18.6, p < 0.0001   R2 = 0.489 (Adjusted = 0.463)
```

Figure 7.11 *ANOVA results when extroversion is added to the model.*

can control for extroversion (i.e., remove the effects of extroversion) by including it in the ANOVA. The ANOVA reports the unique effect of social pressure (i.e., after the effect of extroversion has been taken into account). Whatever remains as the (now) unique effect of social pressure is not contaminated by extroversion and is a purer measure of the effect of social pressure on compliance (see Figure 7.11).

The rows in this ANOVA table correspond to the areas in the Venn diagram in Figure 7.10:

Social pressure	$A1$
Extroversion	$A2$
Error	B
Total	$A1 + A2 + B$

Note that the ANOVA completely leaves out the overlapping area $A12$: the total SumSq has gone down from the previous result (which was 12,832) by an amount that corresponds to $A12$. That part of the variance of the response jointly explained by the two predictors effectively becomes conflated with the unexplained variance. There is a good logic to this – the simplest account of the association between the two predictors is that some unknown variable is behind it.

The Venn diagram wrongly suggests that the unique effect is always smaller than the total effect. This is not true – it is just an artefact of the diagram. For example, imagine our data also had a measure of attitude toward compliance that has a positive effect on compliance. This effect will overlap considerably with the effect of social pressure. The remaining unique effect of attitude on compliance now will, surprisingly, be negative. This is because, when social pressure is controlled for, attitude really represents the extent to which one's attitude is over-ambitious – someone who has higher expectations of what they can do than their social environment will support.

Analyzing Data 4: Inferences about the Population

We must not lose sight of the fact that the best-fitting model is still just a description of the sample and does not yet tell us anything about the population from which the sample was drawn. The final stage in data analysis is to use the (exact) description of the sample to reach an (uncertain) inference about the population.

Concept 4: Uncertainty

The interpretation of data should be seen as a slightly skeptical look at what a sample taken from a population might be telling us about the population itself. Our sample is fixed and our knowledge of the properties of the sample will be exact. However, everyday experience tells us that, had we taken a different sample, its properties would be different. That means that we must have some uncertainty about how much weight to give to our sample. Whatever conclusions we reach about the population must, therefore, allow for the fact that a different sample could have told a different story.

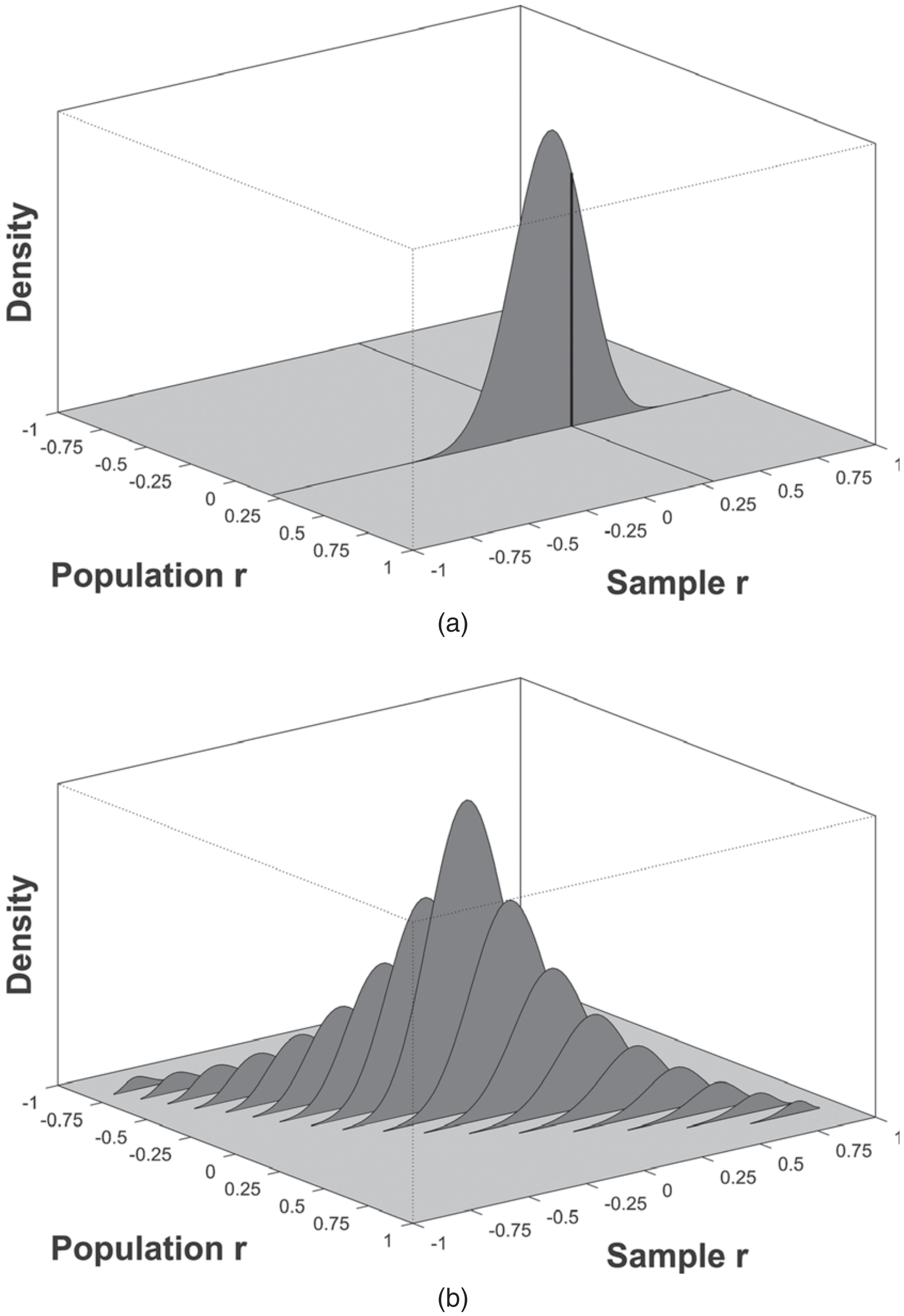
We will use the sample effect size as an estimate of the population effect size. We know that doing this incurs an error, the *sampling error*, which is the difference between the estimate and the true population value. Since we don’t know the population value, we don’t know what our sampling error is. It appears that nothing useful can be said. However, that is to be too pessimistic.

It is possible to be precise about the variability in outcomes from sample to sample for a specific population effect size and specific sample size. Figure 7.12a shows the specific distribution of sample effect sizes that will be produced for a population effect size of 0.2 and a sample size of 42. A researcher who happens to be studying an effect that matches this will get a sample effect size that is drawn at random from that distribution. A distribution like this is called a *sampling distribution*.

Figure 7.12b adds many more sampling distributions for different possible population effect sizes. A researcher has an unknown population effect size, meaning they don’t know which of these sampling distributions their sample belongs to. But they do know their sample effect size, and we can use the same diagram to understand how to use that knowledge to narrow down which population effect size they are studying.

Figure 7.13a takes a specific sample effect size (of 0.3) and shows how relatively frequent this sample is for four different populations (the vertical dark lines). The sample effect size of 0.3 is produced more frequently by the center population than the other two, and we can infer that it is more likely to be the source of the sample than the other two. The sample could have come from any of them, but the central population has the highest likelihood. In Figure 7.13b, we increase the number of possible populations and calculate the likelihood of our sample effect size for each. Eventually, this produces a continuous curve that runs from front to back (Figure 7.13c). This is called the *likelihood function*. Given a sample effect size (of 0.3), it shows the variation in relative likelihood of different population effect sizes given the sample effect size. The highest likelihood is called the *maximum likelihood* and the population that gives it is the *maximum likelihood estimate* of the population effect size.

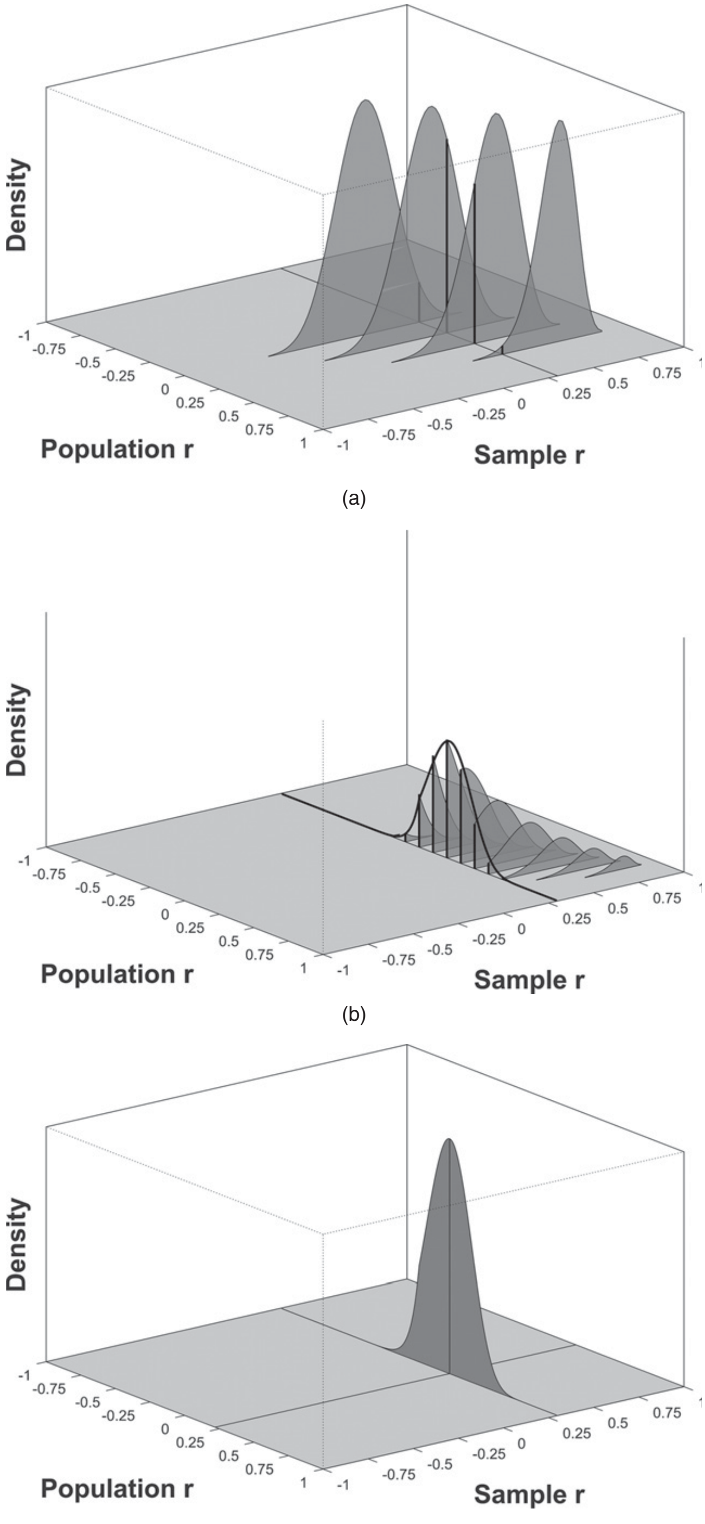
Done this way, the population effect size that has the highest likelihood is the same as the sample effect size. This makes the sample effect size the maximum likelihood estimator for the population effect size. However, we have used a hidden but critical



(a)

(b)

Figure 7.12 *The logic of sampling distributions. Sampling distributions run from left to right. In (a) a sampling distribution is shown: the set of samples that will be obtained from a particular population. In this case it is a population with effect size of 0.2. In (b) many such sampling distributions are shown coming from many populations with different effect sizes.*



(a)

(b)

(c)

Figure 7.13 *We can see the relative likelihood of a sample effect size of 0.3 for (a) three different population effect sizes and (b) many different population effect sizes. (c) The continuous distribution that results when enough population effect sizes are considered. This is called a likelihood function. The population effect size that gives the highest likelihood (the peak of the curve) is then the maximum likelihood estimate of the population effect size for the given sample effect size.*

assumption to reach this point. We have supposed that all population effect sizes are a priori equally likely.

Practical Step 4a: Null Hypothesis Testing

A very specific form of statistical inference that is widespread in social and behavioral science is null hypothesis statistical testing (NHST). The logic of this is simple, slightly counter-intuitive, and driven largely by what was possible with pencil and paper before easy access to computers opened up computational statistics.

NHST asks a clear, unambiguous question of the data – how frequently would the sample effect size or one more extreme occur by chance if there is no effect in the population. The null hypothesis is the hypothesis that the population effect size is zero. The test is to compare what the null hypothesis predicts with what happened. If the comparison suggests an inconsistency, then one or other proposition – the null hypothesis or the data – must be incorrect. The data are not incorrect, and so such an inconsistency, if it occurs, is used to reject the null hypothesis.

Using the sampling distribution for the null hypothesis, we can calculate the probability that the null hypothesis will produce an outcome that is *r* or greater. This probability is called the *p*-value. By convention, if the *p*-value is less than 5%, the result is declared inconsistent with the null hypothesis; this value is called alpha (α). The value for *p* is normally calculated via an intermediate test-statistic, generally *t, F*, or chi-square, with associated degrees of freedom (sample size minus number of model coefficients). Most statistical software calculates *p* (see the final column in the ANOVA tables above). The *p*-value for the data in Figure 7.2 is very small (less than 0.0001) so that result would be considered statistically significant – the null hypothesis that the population has a zero effect size is rejected as inconsistent with the sample.

The *p*-value is simple to visualize in the diagrams we have just used. Figure 7.14 shows the sampling distribution for samples taken from a population with an effect size of zero. The known sample effect size of 0.3 is marked in that diagram. The probability that the null hypothesis will produce this sample effect size or more extreme is then the proportion of the sampling distribution that is beyond 0.3. In this case it is approximately 5%.

NHST tests the null hypothesis by comparing the expected outcomes against the actual data. They either conclude that the null hypothesis and the data are not realistically consistent with each other, and the null hypothesis must be rejected, or they conclude nothing. This is worth emphasizing. If the *p*-value, the probability that the null hypothesis will produce the data or more extreme, is not less than α, then nothing can be inferred. Equally, if the *p*-value is less than α, then the only valid inference is that the null hypothesis is rejected. The alternative hypothesis (that there is an effect in the population) has not been tested and no inference can be reached about it.

With NHST, the presence of uncertainty is hidden in the result. All too often, the outcome – significant or not – is presented without any visible uncertainty. However, the uncertainty is still there. It is the possibility that the inference is wrong

Table 7.1 *Possible inferential errors in the process of NHST*

	$p < 0.05$	$p \geq 0.05$
Effect present in population	Correct inference	Type II error (missed rejection)
Effect not present	Type I error (false rejection)	Correct inference

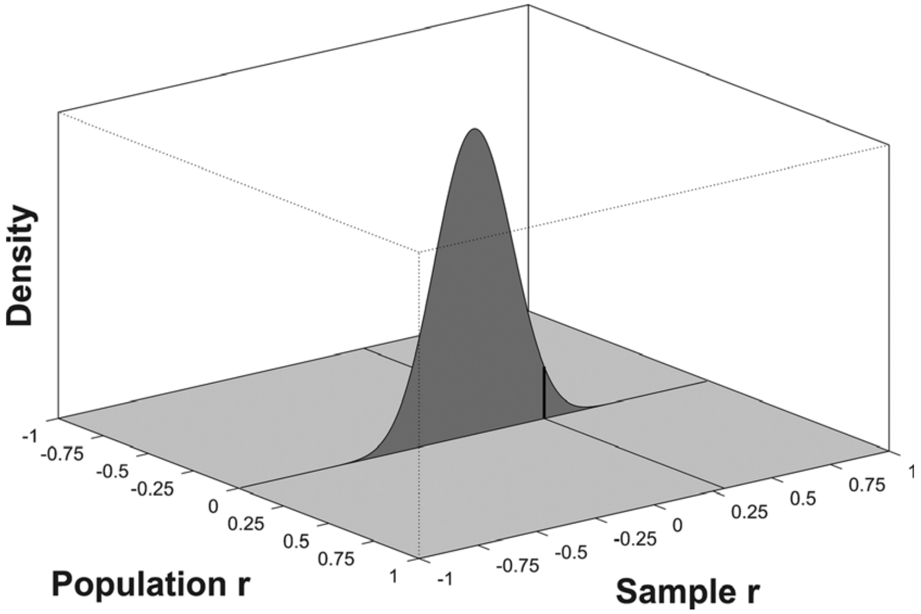


Figure 7.14 *The distribution of expected sample effect sizes from a null hypothesis (population effect size is zero) and a sample size of 42. The proportion of the area under the curve that lies outside of a given sample effect size is the probability that is used for null hypothesis testing.*

(see Table 7.1) in either making a Type I error (false rejection) or a Type II error (missed rejection). Note that these errors are sometimes referred to as false positive or false negative, respectively. However, the concept of a false positive suggests that a positive result has been achieved – the alternative hypothesis has been accepted. This is not what has happened. Worse, the concept of a false negative implies that when p is greater than α, a wrong (negative) conclusion has been reached, but no conclusion should have been reached.

Before a sample is acquired, both types of error (Type I and Type II) are possible:

- For Type I errors, it is possible to say something about how probable it is that an error will be made. In the absence of any knowledge about the population it is either 5% or 0%.

  - ○ If the population effect size is truly zero, then the probability of obtaining a significant result is α, regardless of whether the sample size is five or five million.
  - ○ If the population effect size is not zero, then a Type I error cannot be made.
- For Type II errors, the unknown population effect size and chosen sample size jointly determine the probability of a significant result. Since one of these is unknown, it is not possible to calculate the probability of a Type II error.

After analysis has been done, only one of the two inferential errors is possible, depending on what the outcome is:

- If the outcome is to reject the null hypothesis, then the likelihood that a Type I error has been made is exactly the *p*-value. So, a smaller *p*-value indicates a lower likelihood that a Type I error has been made.
- If the result is not to reject the null hypothesis, then a Type II error may have happened. However, the calculation of the likelihood that this has happened is impossible without the population effect size. Even though the sample provides an estimate of the population effect size, this is already uncertain and cannot safely be used to estimate a further measure of uncertainty.

It is a matter of opinion whether or not NHST is a really productive scientific method. It is certainly weaker than the likelihood approach that we turn to next.

Practical Step 4b: Likelihood of a Population Given a Sample Effect Size

There is a good way of reconceptualizing the idea of best-fitting model. Instead of asking which model (of the data) gives the *best fit* to the data, we ask which model (of the population) is *most likely*, given the data we have. This model is called the *maximum likelihood* model. It is a happy coincidence that, in many practical circumstances, the maximum likelihood model corresponds to the least-squares one. The maximum likelihood model is the best estimate of the population, given the data, but it is important to also find out what the uncertainty is.

The likelihood function (as shown in Figure 7.15) encapsulates all the uncertainty in the sample. It is complete in that sense. However, in an important sense it has too much information. The likelihood function shown in that figure has a non-zero likelihood for an effect size of −0.999 or +0.999. These likelihoods are tiny but not zero. It is probably safe to disregard them, and otherwise, we would be left with the proposition that we can't rule out any possible population effect sizes since they all have a non-zero likelihood. This would represent poor progress for the effort involved. There are various approaches to managing how best to work around this and qualify the maximum likelihood inference.

The most common approach uses the concept of a *confidence interval* (often abbreviated to CI). The confidence interval takes the likelihood function and identifies a band of possible population effect sizes that would encompass 95% of the area under the function. This suggests 95% confidence that the true population effect size lies within that band. We are being careful to talk about confidence not probability

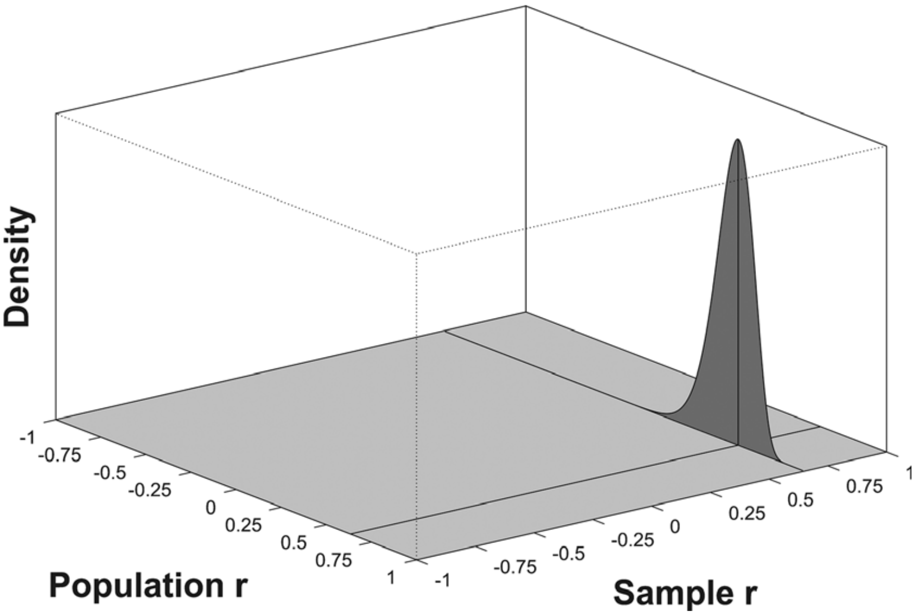


Figure 7.15 *The likelihood function for the data in Figure 7.2. It shows the relative likelihood of different population effect sizes given that sample. The range is quite wide, and the 95% confidence interval is +0.49 to +0.82.*

here. Strictly speaking, the 95% confidence interval will contain the true value on 95% of the times we might do such a calculation. For our sample, we can be 95% confident that the population effect size lies in the range +0.49 to +0.82. The choice of 95% (instead of 75% or whatever) is arbitrary. The confidence interval for a normalized effect size is not symmetric.

A similar approach uses the concept of the *standard error*. Strictly speaking, normally the standard error is defined as the standard deviation of the sample effect sizes produced by a given population (and design). However, the likelihood function can also have a standard deviation which can be thought of as the standard error of measurement for the effect size. Like the confidence interval, this provides a measure of how widely spread are the possible population effect sizes. Since a standard error is calculated as a standard deviation and, therefore, indirectly as a variance, it uses a familiar concept. The data in Figure 7.2 can be described as an effect size of 0.69 ± 0.12.

One could also ask how strongly peaked the likelihood function is. A sharp peak to the function indicates that likelihood falls away rapidly on either side of the peak. For example, we have already seen how an increase in sample size reduces uncertainty.

Finally, we can choose some arbitrary reduction in likelihood (e.g., one-third) and report how far away from the peak the likelihood function crosses that level. This is closely related to how peaked the function is. In Figure 7.16, the function drops to

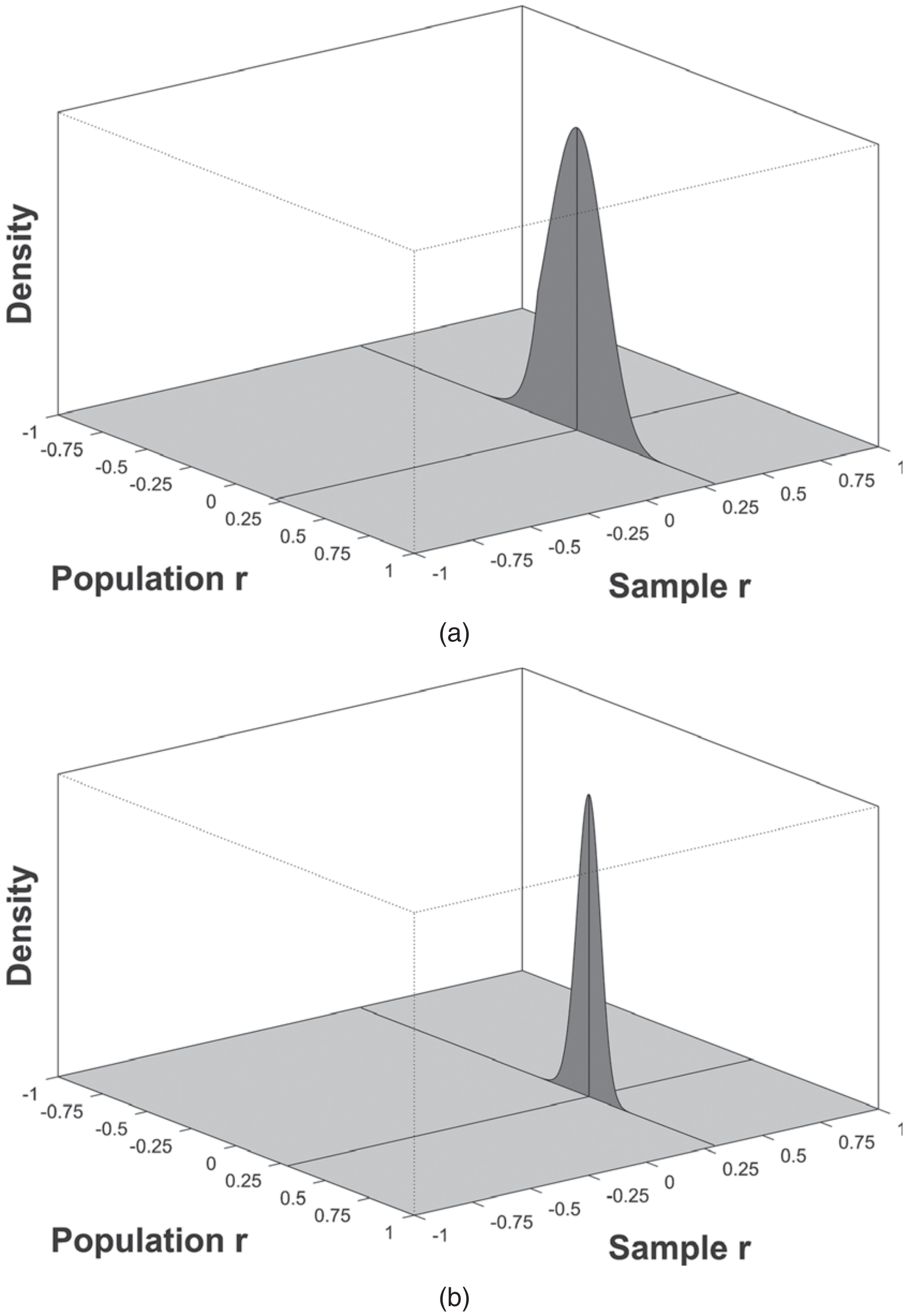


Figure 7.16 *Two different likelihood functions are shown for two different sample sizes. (a) The smaller sample size (*n *= 42) leads to more uncertainty than (b) the larger sample size (*n *= 500). Reflecting this, the second likelihood function has a much sharper peak.*

one-third of the maximum over a broader range of population effect sizes for $n = 42$ (Figure 7.16a) than for $n = 500$ (Figure 7.16b).

Reaching Conclusions

The purpose of data analysis is to allow us to reach conclusions. But it must never be forgotten that conclusions about the population from a sample are tentative and provisional. The uncertainty that begins with how the random sample relates to the population remains at the end of the process; there is that same uncertainty in any conclusion about the population.

Which Population?

The first question to consider before reaching a conclusion is about the population itself: What is the population that this sample can be said to come from? We wish to use our inference to generalize to everyone who could have been in our sample.

There is a technique of data analysis called bootstrap that helps to understand this. Essentially, bootstrap builds an artificial population that is created with uncountable replicates of each participant in the sample. It then takes a sample at random from this population and analyzes that. The random sample will have repetitions of some participants, and others will be missing. This process is repeated many times to give the sampling distribution for the process. This procedure makes explicit what the population is. The lesson is that it is sensible to think of the population as being just the endless repetition of the sample. If the sample is a set of first-year psychology students, the population is an unlimited supply of similar first-year psychology students.

If our sample is limited in any way, it is important to reflect that limitation in any conclusion reached. Of particular issue in this respect is how our sample is drawn from the population. If the sample are all volunteer participants, then the sample is only safely valid as a sample of the population of people who will volunteer.

How Much Uncertainty?

The inference about a possible population from a sample is uncertain, and this has to be at the heart of any conclusion that is reached. Any conclusion is provisional – to be refined by further samples. In this section, ways of assessing the uncertainty are explained.

Often, a conclusion reached from analysis of a data set is a binary decision – the analysis shows or doesn’t show a relationship between two or more variables. This is frequently the result of hypothesis testing, usually null hypothesis testing, although other forms of hypothesis testing (within the Bayes framework, for example – see Chapter 23 of this volume) lead to the same outcome. When this logic is used, it must be adhered to; the only valid outcomes are that either a hypothesis is rejected or nothing is concluded. Note that conclusions that a hypothesis is rejected or not convey a degree of finality that is simply not merited by the procedure itself.

In this binary conclusion, the uncertainty appears to have vanished, but it has not. It remains in the possibility that the inference is incorrect.

Nonetheless, hypothesis testing is frequently interpreted as showing that an effect exists or that it does not exist. This inevitably leads to a situation where two perfectly valid samples can lead to apparently contradictory conclusions. It is only by going back to the uncertainty in each sample that this can be resolved. There is a modification to this procedure that goes some way to dealing with this and is commonly associated with a Bayesian philosophy. Before analyzing data, there is some starting confidence that the effect in question does or does not exist in the population. Then, the analysis of the data proceeds and either increases our confidence in the hypothesis or weakens it. This is an important step forward – the concept of confidence incorporates the uncertainty.

If the likelihood approach to inference has been adopted, to a degree uncertainty is already present. To say that 0.234 is the most likely population effect size is already to acknowledge that there are other, albeit less likely, possibilities.

Causation?

It is often said that "correlation does not mean causation" but this is not quite right. It is better to say that the existence of a relationship between two variables, *a* and *b*, probably does have a cause, but we cannot say which variable causes which. Given a relationship between two variables, only one of these three statements is correct:

- *a* causes *b*
- *b* causes *a*
- *a* and *b* are both caused by some further variable *c*

However, if we know completely how one of the variables, *a*, was caused and that its cause was entirely independent of *b*, then we can rule out the second and third statements and then safely infer that the remaining first statement *a* causes *b* must be true. This often happens when *a* is an experimental variable whose value is assigned by the experimenter to each participant. When a participant is randomly assigned to either an active treatment or a non-treatment group, for example, the experimenter is causing the treatment variable. So long as the assignment is made in a way that doesn't depend on the participant or their outcome, then a causal conclusion is fair. However, if the ultimate assignment of a participant to one of the groups depends on the participant (e.g., their success in complying with the treatment), the causal conclusion is no longer fair. We cannot move participants who we assigned to the active group into the non-treatment group because they didn't complete the treatment. Doing that means that the cause of group membership is no longer completely known.

Which Effects?

Null hypothesis testing is a specific form of a typical question that is asked of data – does an effect exist? The answer usually depends on the effect being big enough

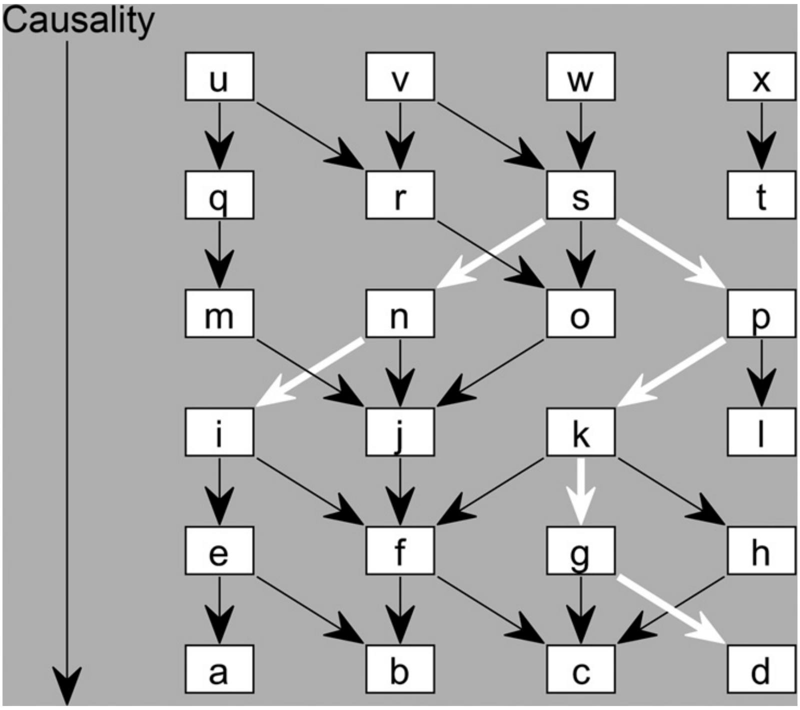


Figure 7.17 *Each rectangle is a variable, and the arrows are causal links. In this observational system, there will be a small statistical effect between* i *and* d *that is not causal (i.e.,* i *and* d *are both caused by* s *and its causes – the white links). There is also an effect of* i *on* c*, part of which is causal (via* f*) and part of which is not (via* n *and* j *and via* s*,* p*, and* k*).*

to detect. There is reason for considering the possibility that effects always or nearly always do exist but may often be very small.

Most systems that underlie research in social and behavioral sciences are complex, with many different influences working together. Figure 7.17 shows a network of 24 variables with a modest degree of complexity, which is probably a reasonable mental model for the systems in science. It is quite sparse – only 5% of the 500+ possible links between the 24 variables are present. Despite being relatively sparse, there are very few pairs of variables that are not connected indirectly. Indirect connections can still lead to associations. For example, values at *n* and *o* are associated because variable *s* drives them both – the classic case of a confounding variable. In fact, the variables that *n* is not linked to and therefore not associated with are just four: *u*, *r*, *q*, and *m*. The twofold rule here is that: (i) you can go down the diagram following any links; (ii) you can start by going up the diagram, but you can only change direction to going down once. The total effect measured between two variables is the sum of the effects of all the different paths between them.

It is rare for a pair of variables to be entirely unrelated (e.g., *u* and *d*). In such a network, the most useful question to ask of data is not whether the effect exists, but how large it is. In this network, there is an effect linking *i* and *d* and an effect linking *k* and *d*. Both effects exist and, given a large enough sample, each could be demonstrated. They will be of very different effect sizes, however. If the typical normalized effect size for each link is 0.5, then the effect size for *i* to *d* would be $0.5^6 = 0.016$; the effect size for *k* to *d* would be 0.25, which is 15 times larger.

Interestingly, the only route from *i* to *d* goes through *k*. This means if one were to measure those three variables, *i*, *k*, and *d*, the effect sizes linking them all to each other will show a pattern that indicates, subject to sampling error, there is no link from *i* to *d* that doesn't go through *k*. The pattern is simple:

$$r\ (i \rightarrow d) = r\ (i \rightarrow k) \times r(k \rightarrow d)$$

This would be described as a *mediation analysis*. The effect of *i* on *d* is mediated by *k*.

The moral here is that the approach of asking whether an effect exists is typically inadequate for understanding complex systems. Only by systematically obtaining estimates of the effect sizes can that be undertaken.

Conclusion

In this chapter, we have sought to give the reader an orientation to statistics. We have focused on three core issues in data analysis:

(1) **What is the population that the data represents, and how close is the sample to the ideal of random sampling?** The answer determines how safely we can generalize our results from the sample to the intended population and how much uncertainty we should retain in doing so.

(2) **How do we describe the patterns that are to be found in the sample?** The answer is that, by fitting a model to the data, we can both describe the patterns and quantify them as effect sizes.

(3) **Where is the uncertainty in the conclusions that we reach?** When presenting the results of data analysis, the reader must be told how much uncertainty there is.

The researcher who ritualistically jumps straight from a spreadsheet to a *t*-test, only seeing the *p*-value, has missed out on something important by not considering these questions.

Further Reading

Most of the material this chapter is common to all recent accounts of statistical analysis in the social sciences. We therefore have not provided references in the text. However, an interested reader will wish for pointers to more reading. The following texts are recommended for more in-depth accounts of many of the topics covered:

Barford, N.C. (1985). *Experimental Measurements: Precision, Error and Truth*, 2nd ed. Wiley.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioural Sciences*. Psychology Press.

Cumming, G. (2012). *Understanding the New Statistics*. Routledge.

Edlund, J & Nichols, A.L (2019). *Advanced Research Methods for the Social and Behavioural Sciences*. Cambridge University Press.

Efron, B. & Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press.

Ellis, P. D. (2010). *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge University Press.

Hand, D. J. (2008). *Statistics: A Very Short Introduction*. Oxford University Press.

Harlow, L. L., Mulaik, S. A., & Stegier, J. H. (eds.) (2016). *What If There Were No Significance Testing?* Routledge.

Hayes, A. F. (2013). *Mediation, Moderation and Conditional Process Analysis*. Guilford Press.

Lambert, B. (2018). *A Student's Guide to Bayesian Statistics*. Sage.

Pawitan, Y. (2013). *In All Likelihood*. Oxford University Press.

Pearl, J. (2009). Causality: Models, Reasoning and Inference, 2nd ed. Cambridge University Press.

Rosenthal, R. & Rubin, D. B. (1982). A simple general purpose display of magnitude and experimental effect. *Journal of Educational Psychology*, *74*, 166–169.

Rosnow, R. L. & Rosenthal, R. (1997). *People Studying People: Artefacts and Ethics in Behavioural Research*. Freeman and Co.

Watt, R. J. & Collins, E. C. (2019). *Statistics for Psychology: A Guide for Beginners*. Sage.

Published online by Cambridge University Press

8 Writing the Paper

John F. Dovidio

Abstract

Writing the paper is one of the most challenging aspects of a project, and learning to write the report well is one of the most important skills to master for the success of the project and for sustaining a scholarly career. This chapter discusses challenges in writing and ways to overcome these challenges in the process of writing papers in the social and behavioral sciences. Two main principles emphasized are that writing is (a) a skill and (b) a form of communication. Skills are developed through instruction, modeling, and practice. In terms of communication, the research report can be conceived as a narrative that tells a story. Sections of the chapter focus on identifying common barriers to writing and ways to overcome them, developing a coherent and appropriate storyline, understanding the essential elements of a research paper, and valuing and incorporating feedback.

Keywords: Ethics, HARKing, Hypothesis Testing, Learning Mindset, Persuasion, Scientific Writing, Writer's Block

Introduction

Completing the task of writing the paper is a necessary element of a successful scholarly career in the social and behavioral sciences. In the assessment of a scholarly record in many behavioral science fields, the peer-reviewed research article is the primary element for evaluation. The centrality of writing to an academic career is often referred to, somewhat pejoratively, as "publish or perish". However, for most professionals and students, publishing is motivated intrinsically by a desire to share one's finding with others. Also, from the perspective of the field, scientific discovery is primarily valuable when findings are communicated broadly. Different disciplines vary in research report format, length limits, and writing conventions. However, all disciplines value original scholarship that is communicated professionally, clearly, and persuasively. Thus, this chapter is not just about writing the paper; but it is also about doing so effectively.

The quality of writing is important for many reasons. As evidenced historically in formal reviews of books and currently in the form of blogs and other types of social media, how well a piece is written determines how many people will choose to read it and what the response to the message will be. Although a professional audience is typically a more captive one than popular audiences because scholars need to stay current with the scholarly literature on a particular topic, the quality of writing still plays a critical role.

Writing quality is influential, for example, in the scholarly peer-review process. Even though the primary criterion of peer review is the assessment of scientific merit, the effectiveness of the writing of the report affects the overall reviewer evaluation directly and indirectly. In an analysis of outcomes of manuscripts submitted to a leading psychology journal, clarity of presentation was the third strongest predictor of the favorability of a reviewer's recommendation, only ranked behind significance of the contribution and rigor of the methodology (Dovidio, 2010). Manuscripts that were recommended for acceptance or revision had an average rating of a 4.2 on a 1 (poor) to 5 (excellent) scale referring to the clarity of presentation; manuscripts that were not recommended had a significantly lower average rating of 3.2.

Even when not identified as an explicit criterion for review, the quality of writing shapes reviewers' and editors' impressions of a manuscript. Research on psychological fluency reveals that people think more deeply about and find written material more persuasive when it is presented in a way that is easier to process (Oppenheimer, 2008). Yet, writing receives much less explicit instruction in professional training curricula than do other key scholarly competencies – certainly much less than research design and statistics and typically much less than teaching.

This chapter discusses writing the paper with a broad perspective. It considers the mechanics of writing a report in the social and behavioral sciences, but it also examines the process of writing from preparing to revising. It emphasizes a general approach to writing and presents guidance from leading scholars. Admittedly, because of my own professional training and experience, this chapter has a psychological bent and an emphasis on reports of experimental studies. However, the aim is to consider topics, challenges, and suggestions that apply broadly across disciplines and research paradigms in the social and behavioral sciences. The first section (Approaching the Task) briefly highlights some common barriers to writing and offers suggestions for overcoming them. The second section (Developing the Story) emphasizes the value of understanding the main message that is intended to be conveyed before one begins to write. The third section (Creating the Narrative) reviews the key sections of a research paper and offers suggestions for communicating the material effectively. The final section (Revising the Report) highlights the value of seeking and receiving guidance for changes and offers recommendations for responding to reviews.

Approaching the Task

There are several excellent books offering guidance in scholarly writing (Baglione, 2020; Becker, 2020; Rocco & Hatcher, 2011; Sternberg & Sternberg, 2010). However, prerequisite to writing effectively is the task of sitting down to write *something*. This section is about preparing oneself to begin the often-arduous task of writing. It considers why people frequently experience obstacles as they prepare to write, particularly with respect to scholarly writing, and discusses strategies that help to facilitate writing.

Despite widespread understanding of the central role of publication for a scholarly career, many scholars describe writing the report as one of the most challenging stages in the research process. In fact, many highly skilled and brilliant scholars have left the field, either voluntarily or involuntarily, because of their inability to write the research paper. There are several well-documented reasons why people experience "writer's block" – the inability to produce new material. General fear of evaluation or rejection (Boice, 1993) and situation-relevant stressors create excessive arousal and inhibit the creative processes that are necessary for effective writing (Byron et al., 2010).

Whereas developing and testing a research idea are relatively private activities, it is the written report that puts the research under a more public microscope as it enters the peer-review process. Seeking publication of a research report through peer review is an essential scholarly pursuit, but it is an activity that people typically engage in with trepidation. The peril that people experience when they must produce a paper that will be scrutinized by others is not necessarily misguided. People experience a sense of ownership over their ideas and the products of those ideas, and thinking about them activates parts of the brain that are similar to when people think of the self (Morewedge & Giblin, 2015). Obviously, people do not like rejection, and this applies to the products of their ideas – such as a research paper. Rejection of this extension of the self activates a range of neural, hormonal, and psychological processes reflective of threat reactions (Kim & Johnson, 2015). Yet, prestigious journals often have rejection rates over 85%. Thus, even factoring in the characteristic tendency to underestimate the likelihood that one's own manuscript will be rejected (Moore & Schatz, 2017), the possibility of rejection still looms large for researchers. Every word that is written puts the person one step closer to the peer-review stage in which rejection is most likely to occur. Not writing, while counter-productive in the long term, can alleviate fear and anxiety in the short run.

There are several strategies directed at relieving stress and anxiety to overcome writing paralysis. These include taking a break from writing, ameliorating tension with activities such as exercise, and strengthening emotional and cognitive resources by getting an appropriate amount of sleep and eating healthy foods. Because writer's block occurs when people have lost confidence in their ability to write (Boice, 1993), engaging in activities that restore one's sense of mastery and bolster self-confidence is another effective way to combat writer's block. The anxiety associated with writing can be reduced, and confidence in writing can be restored, by engaging in freewriting, in which people write for a specified amount of time, often recommended to be about 15 minutes, without worrying about grammar or rhetorical conventions.

The debilitating fear and anxiety that people experience as they approach writing the report are exacerbated when researchers hold unrealistic standards for themselves. Being a perfectionist is counterproductive for scholarly writing. The perfect paper has yet to be written, and setting the goal of writing the perfect paper can be paralyzing. Also, when perfection is the standard, even the most helpful constructive comments from others are wounding because they represent failures. Whether thinking of oneself as a perfectionist precedes efforts to write the report or occurs to justify lack of progress in writing, it does not contribute positively to the task at hand. High quality is the appropriate objective; the goal of perfection is unattainable and thus often debilitating.

Even among scientists who may not experience the extreme debilitating effects of fear of rejection, the writing phase of a research project is particularly challenging. People who pursue a career in research tend to be intellectually curious and motivated by a desire for discovery, both in terms of general principles and with respect to insights into issues of personal relevance (often called “me-search”). They are energized by the opportunity to pursue a meaningful question, stimulated by the challenges of developing ways to test these ideas, and intrigued by deciphering the data collected to address the question. But once the creative stages of formulating the research question, creating the methodology, and analyzing the data have been completed – steps that are typically performed before any writing of the report begins – the researcher’s motivation wanes. Learning the answer to a professionally and/or personally important question and communicating the finding to others are fundamentally different enterprises. The transition from doing the research to writing up the research is thus a particularly precarious one.

When anxiety rooted in fear of rejection or evaluation or created externally by impending deadlines cannot be eliminated, people can still overcome writing paralysis by creating support structures. Writing groups that can assemble in-person or rely on computer-mediated communication platforms (e.g., Zoom) provide social accountability and encouragement among people in similar situations (Chai et al., 2019). In addition, writing can be thought of as an exercise. People are more likely to continue to follow an exercise regimen when it is a regularly scheduled activity rather than performed at various times of convenience. Individuals differ substantially in the times of day they are most alert, generative, and efficient, and these scheduled writing periods need to be situated within these highly productive time frames. Also, these writing sessions need to be free of distractions (e.g., emails, text, and messages) and avoidant actions (even when they can be rationalized as relevant, such as searching for new scholarly sources). In addition, like exercise, as people build up writing stamina, these sessions should be made longer and more demanding. Also, in collaborative projects, when one person stops making consistent progress in writing, a co-author should be prepared to take on the responsibility of writing. Such a tag-team approach ensures continuous progress and allows the author whose efforts have stalled to engage in activities used to overcome writer’s block.

Although people commonly refer to an individual’s *talent* for writing, writing is primarily a skill – a learned ability. Like any skill, some people are better than others; however, anyone with the appropriate background, knowledge, training, and effort can become sufficiently proficient at writing to have a successful scholarly career. The basic principles of skill acquisition and learning generally – engaging in the activity repeatedly over time, receiving and appropriately responding to feedback, and reinforcing good habits (Wood, 2019) – apply to developing strong writing skills. In addition, like any exercise activity, the more you write, the easier it becomes. While always a challenge, with practice and experience writing becomes less stressful and even enjoyable.

In summary, there are many reasons why it is difficult to begin and sustain efforts to write the report. Nevertheless, it is important to be an effective “finisher” as well as being good at conducting research. If you have seven reports that are 90% done, you have no full manuscripts. Only completed manuscripts can be submitted for publication. Although the obstacles to writing are many and complex, there are several well-established techniques for overcoming writing paralysis. However, the challenge is not just writing the specific report at hand but also developing generalizable skills to writing research reports across a career. The sections that follow are intended to provide guidance in developing and honing those skills.

Developing the Story

Communication is not just about what information is conveyed by the source – in this case, the researcher – but also by how that information is received by the audience. People often think of scholarly writing as distinct from popular forms of writing. Consequently, they feel that it does not have to be engaging, well-written, or even coherent. They are wrong. As explained earlier, the quality of writing affects how a manuscript will be reviewed and how successfully it will attract an audience within and outside the profession. This section offers guidance in how to (a) organize your thoughts, (b) envision a story that represents the ideas and hypotheses that originally guided the work, (c) be focused and direct in structuring the story, and (d) tailor the story for the intended audience.

Understand Where You Want to Go

One of the most basic elements in preparing to write is to formulate the story you want to tell. However, to overcome writing inertia, people often focus on starting to write but without a clear story in mind. While this helps alleviate anxiety in the short run, it often leads the writer – and eventually the reader – far astray from the central message that needs to be conveyed when writing effectively. The process of developing your story is thus a critical first step.

In this process, begin at the end. What point do you want to make? What is the conclusion you want to draw? Then, outline the steps in the argument and evidence that lead to that conclusion. That is, know where you want to go before you go there. Among the many articles and chapters written about scholarly writing, one of the most common foundational recommendations in preparing to write is to understand the story you want to tell before you start writing. Roediger (2007) advises, “Provide an easily remembered take-home message. You should provide clear answers to the following two questions the reader will have: What has the paper told me that I did not know before? And why is this news important?” It is critical to understand that just because you devoted so much time and effort to study a topic that is fascinating to you does not mean that the research you conducted is either interesting or important to others. The burden is on you to make that case; you need to sell your ideas actively.

Be True to Yourself (and Others)

The story you tell also needs to reflect the ideas and logic that guided the project; in particular, it should be faithful to the hypotheses you developed before you embarked on data collection. It is important to note that the conventions and standards for writing research papers sometimes change over time as the field becomes aware of the negative consequences of some common practices. For instance, for many years (through the 1990s), behavioral science researchers were commonly advised to write "the article that makes the most sense now that you have seen the results" (Bem, 1987, p. 172). That advice is outdated, misguided, and clearly at odds with present ethical standards for research (as acknowledged more recently by Bem, 2000, 2004). It reflects the currently discredited practice of HARKing (hypothesizing after the results are known; Kerr, 1998).

There are two main problems with the practice of creating a narrative based on findings rather than original hypotheses (i.e., HARKing). First, it misrepresents the foundational ideas for the project. Scientific writing, at its essence, needs to be non-fiction. Another problem is the practical impact on the validity of the conclusions drawn. Because of the assumptions of inferential statistics, HARKing inflates the likelihood of making a Type I statistical error. A Type I error is a "false positive", an erroneous conclusion that there is a significant effect in the study (and with HARKing, one that appears to support the researcher's hypothesis) when the result instead occurred by chance. A story that is rewritten in a way that portrays an unexpected significant effect as a predicted significant finding misleads other scholars and readers about the validity, and thus the replicability (reproducibility), of the effect. Indeed, the practice of HARKing was one of the factors that produced the "replicability crisis" in social psychology (Simmons et al., 2011), in which both researchers and the general public became skeptical about findings in the social psychological research literature (Nelson et al., 2019). So, the story you tell should be an honest one; reflect the literature, theory, and evidence that guided your work; be a fair representation the data; and present conclusions based on the full findings.

It is also important to keep abreast of contemporary best practices about other research standards at earlier stages of the project. Today, a substantial number of journals encourage or require preregistration of hypotheses, measures, and intended analyses prior to data collection to limit the number of "false-positives" (Type I errors) in the literature (Nosek et al., 2019).

Stay Focused

Even within the constraints of being faithful to the original hypotheses, there is considerable freedom in how you develop your narrative. In thinking about the story that you want to tell, remember that more is not necessarily better. Unnecessary complexity can obscure the take-home message and make the story less compelling. Also, presenting more studies in a manuscript does not necessarily make a manuscript stronger in the eyes of editors and reviewers. Their evaluations are not an additive model in which adding a study of limited value increases the favorability of the evaluation. Instead, editors' and reviewers' impressions tend to reflect a weighted average, in which they weigh the weakest element of a package of studies most strongly in making their final recommendation (Dovidio, 2010). Thus, keep the story crisp, clear, linear, and strong. There should be no surprises and no tangents in scholarly writing.

Meet the Expectations and Needs of the Audience

One common piece of advice for effective writing is to know your audience and tailor the story to that audience. There are at least four types of audiences to consider: (a) the journal, (b) the review team (the editor and peer reviewers), (c) the profession, and (d) the general public. Even within the same discipline or subdiscipline, journals have different missions, guidelines, and standards. Missions vary along several dimensions, such as emphasis on topic (e.g., of interest across disciplines, within a discipline, or within a subdiscipline), type of contribution (e.g., theoretical and/or applied; empirical vs. review); and methodological focus (e.g., quantitative and/or qualitative). Guidelines for length can differ dramatically (some with limits of 1,000 words; many with 3,000–5,000-word limits; and several with no limits), and journals use a variety of heading structures and referencing styles. Submissions that do not align with the mission or violate the guidelines of a journal are often "desk-rejected" (i.e., rejected without additional reviews) by the editor. The key point here is that authors should decide on the target journal before they start writing.

The editors and peer reviewers – the second audience to consider in preparing to write the report – are in critical gate-keeper roles in determining whether one's work will be published by the journal. This review process is usually very rigorous and produces high rejection rates. In fact, when submitting a manuscript for publication, it is unrealistic to aim for acceptance. In my experiences as an author, reviewer, editor, and publication board member across over 40 years, I have not personally been involved in a review process in which a manuscript was accepted without revision. The best realistic hope for authors is that their manuscript will be invited for revision. A revised version that is responsive to the editor's and reviewers' comments will have a reasonable chance of eventual acceptance.

Reviewers have a substantial influence on what gets published and what does not; that is their key role (see Chapter 33 in this volume). Reviewers tend to approach manuscripts with a rejection mindset that derives from information about the generally high rejection rates of journals and creates a norm of rejection among reviewers. Another reason why reviewers tend to recommend rejection of manuscripts is that people who express negative, compared to positive, evaluations of others' intellectual products are perceived to be more intelligent, competent, and expert – qualities that have valuable reputational benefits in academia – even when the work is objectively of high quality (Amabile, 1983).

In part because people tend to seek evidence that confirms their expectations (a confirmatory bias), there are only a limited number of problems that reviewers need to detect before they conclude that the manuscript should be rejected (Garcia et al., 2020). I have used the metaphor of "five gold coins" to describe this process. The metaphor is meant to be illustrative, not factual. Five is an estimate based on experience, not an empirically verified quantity. The assumption is that authors have five gold coins as currency of credibility with reviewers. Every time a reviewer questions a point, pauses because of doubt or confusion, discovers an over- or misstatement, or detects an error in logic, methodology, or statistics, the author loses at least one gold coin. After the fifth gold coin has been expended, the reviewer's mind has been made up and rejection is recommended. Authors can tell when they lost their last coin from the tempo of the written review. The detailed points made by the reviewer become briefer and more general when that last gold coin has been spent. These gold coins of good will and credibility are a precious commodity for authors, and they should not be squandered on confusions, questions, or suspicions created by writing of poor quality.

The third audience to consider is the one composed of potential readers of the work once it is published by the journal. A successful scholarly career is not simply based on how often people publish high-quality work; it is also based on the scholarly impact of the research. One common metric of scholarly impact is the frequency with which research is cited in other scholarly works. This metric applies not only to the assessment of a particular paper but also to the evaluation of researchers themselves, in terms of a person's *h*-index. The *h*-index reflects how much a person publishes in relation to how often those works are cited in other papers. To be cited by others, the research needs to be of high quality and written in a way that attracts and engages a broad scholarly audience.

The fourth relevant audience is the general population. For most scholars, research is not published solely for the sake of research; it is motivated by a desire to provide knowledge and information that will improve the lives of individuals and society (Hawkins et al., 2007). Moreover, through internet resources and social media, research is now more widely accessible to the public than ever before. While communicating with a broad, lay audience is different than speaking to specialists, the core principles of effective and engaging writing are, fortunately, similar: "The story opens with the problem to be solved (a mystery), foreshadows how the research speaks to the problem, and highlights evidence that, when taken together, presents a coherent picture that helps move toward a solution" (Dovidio & Gaertner, 2007, p. 105).

For all four audiences, clarity is paramount. As Becker (2008, p. 412) observed with respect to writing in sociology, "Clarity and precision aren't complicated requirements, but they're not nothing. It takes a lot of care and some skill – not an enormous amount, but some – to put together sentences, paragraphs, and chapters whose point a reader won't misunderstand. To do that, we have to define our terms carefully and make our concepts clear". The next section describes ways to write the report clearly and effectively to attract a wide readership and provide an attractive vessel for communicating the work to others.

Creating the Narrative

The main narrative of the report is typically structured around four sections: (a) introduction, (b) method, (c) results, and (d) discussion. These sections are preceded in the report by an abstract that summarizes the report and is made freely accessible. In this part of the chapter, I discuss materials that are contained in these sections, as well as the abstract, and how that information can be conveyed effectively. I also draw upon the wisdom of distinguished scholars who have offered their own advice.

One general bit of advice about the process of writing the narrative, which applies to each of the report sections, is to begin by writing the best paper that you can. Say what needs to be said and then worry about trimming words. Trimming words is a painful experience, and if you try to cut words too early in the writing process, you will become distracted at a time when you need to be generative in your thinking (rather than constrictive in your prose). Cutting prematurely can interrupt your writing momentum.

It is also important to be careful and attend to detail in your writing. Cues about the general and professional competence of the writer that are embedded in the way the paper is written also affect the receptivity of the various audiences to your message. The recognition that “peripheral cues” – cues that are not directly related to the strength of ideas and arguments in a message (Petty & Brinol, 2011) – influence how the research will be judged led to masked review journal policies. Empirical research has revealed that status-related characteristics of authors (e.g., in terms of the prestige of their institution, standing in the field, and demographic characteristics) systematically and often unfairly influenced evaluations of the work (Lee et al., 2013).

Grammar, spelling, and writing style are key peripheral cues embedded in the manuscript. The degree to which the writing conforms to general grammatical conventions is an influential peripheral cue because it is perceived to convey information about the intellectual competence of the author. Writing in a way that violates current standards for grammar and spelling is especially damaging for an author because grammar and spelling checking functions are widely available. Thus, having errors in grammar and typos is commonly attributed to intellectual carelessness and undermines the credibility of the author and report. It is important to recognize, though, that rules for grammar and style evolve over time. For instance, splitting an infinitive is now acceptable among grammar experts. The word “since” is currently confined to temporal relations; “because” signifies a causal connection. In terms of style, the active voice is now strongly preferred over the passive voice.

Beyond these general standards, disciplinary organizations often specify particular rules, such as whether to hyphenate a term or not (e.g., “inter-group” or “intergroup”) or permitting the use of “they” to refer to a single individual to avoid binary-gender references. Journals also differ in the ways they want authors to refer to gender identity, sexual orientation, race and ethnicity, body size, and disability status. It is important to familiarize yourself with the conventions used in the discipline and specific journal in which you aspire to publish.

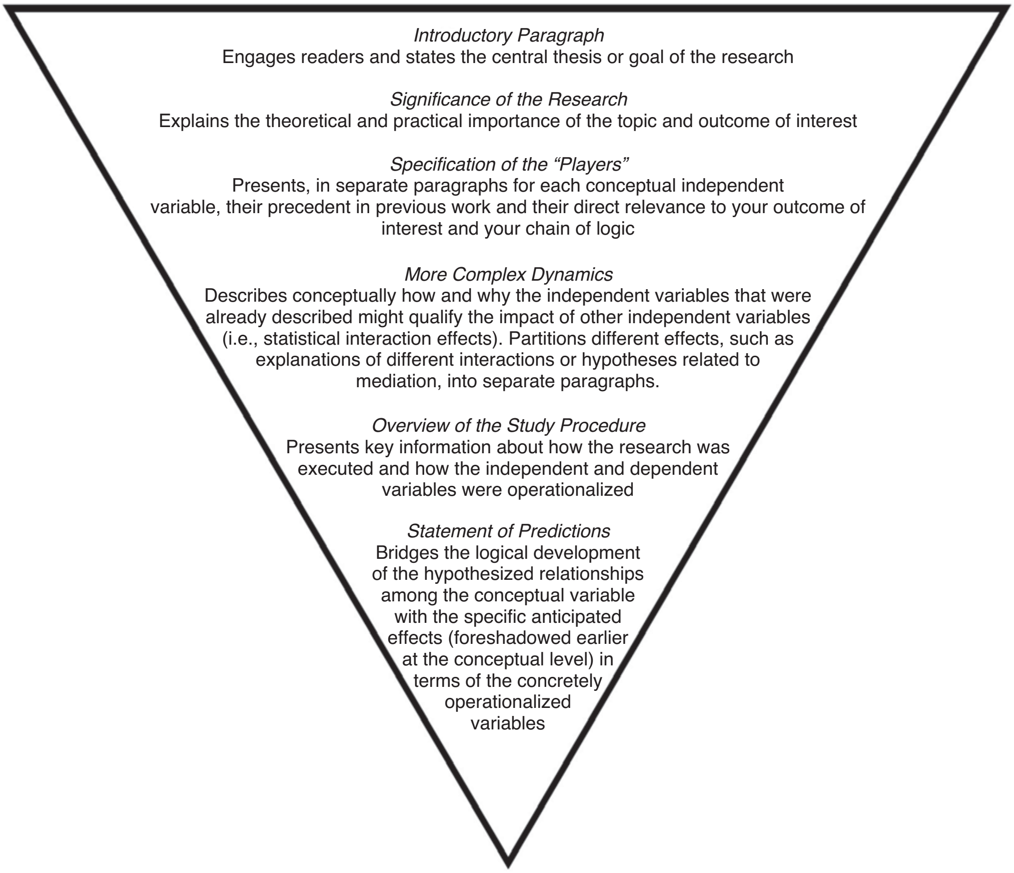
Introduction

Although the general arc of the storyline of the research should be developed before initiating writing, a more specific outline should be created for each major section of the report. The introduction of the manuscript is a form of expository, or persuasive, writing. The objective of expository writing is to convince the reader of the validity of a particular position. At the forefront of an expository paper, in the introductory section, should be a direct statement of thesis that identifies the topic or problem you are addressing and explains the purpose of the project with respect to how it changes the way people think about the topic or addresses the problem. The statement of thesis prepares readers for argument that you will be making. The objective of the paper is then to make a persuasive case that you have the answer for the problem. In the case of the introduction of a manuscript, the crux of your argument is represented in the main predictions. Readers need to be persuaded that the predictions make sense and offer an important and novel contribution to the literature, even before they read the hypotheses. This is done most effectively by having a direct chain of logic that is supported by evidence, while deftly anticipating alternative answers and convincing readers that your answer is the best, correct one. In an empirical paper, the evidence is the support of previous research for each of the key steps in the argument you are making in the introduction, coupled with the new data and analysis you are presenting.

The conceptual structure of the introduction, as in most expository writing, should be an inverted pyramid. The vertex at the bottom of the triangle is the main prediction. Take, for example, a study that investigates how the gendered nature of a topic (traditionally masculine, traditionally feminine, or gender-neutral) affects the amount of time that male and female participants speak during an interaction. The prediction is that when the topic is of a more masculine or gender-neutral nature, male participants will talk more than female participants; however, when the topic is a traditionally feminine one, female participants will talk more than male participants. Figure 8.1 presents basic components of the introduction within the inverted pyramid structure for this study. Each component represents at least one paragraph in this section.

As with writing in general, each paragraph should begin with a topic sentence that succinctly states the main theme of the paragraph and end with a concluding sentence that summarizes and helps make the transition to the next paragraph. No paragraph should be longer than a page; readers' attention will wane before they get to the end. Very short paragraphs may also be problematic. While a short, three-sentence paragraph is easy to read, it is unlikely that, counting the topic and concluding sentences as two of the three, you can present a persuasive rationale in a single sentence. And, if all that can be said about the topic is three sentences, it may not be important enough to include. Finally, it is helpful when you create your outline to also include the topic sentence for each paragraph to get a sense of the logic and flow of the argument you are developing before you begin writing.

The *introductory paragraph* of the paper should begin broadly by capturing the interest of readers. Sternberg (1993) recommends, "Start strong" – research revealed "that 83% of readers never got beyond the first paragraph of the majority of articles they began to read". In the example I presented, the first paragraph might begin by



Introductory Paragraph
Engages readers and states the central thesis or goal of the research

Significance of the Research
Explains the theoretical and practical importance of the topic and outcome of interest

Specification of the "Players"
Presents, in separate paragraphs for each conceptual independent variable, their precedent in previous work and their direct relevance to your outcome of interest and your chain of logic

More Complex Dynamics
Describes conceptually how and why the independent variables that were already described might qualify the impact of other independent variables (i.e., statistical interaction effects). Partitions different effects, such as explanations of different interactions or hypotheses related to mediation, into separate paragraphs.

Overview of the Study Procedure
Presents key information about how the research was executed and how the independent and dependent variables were operationalized

Statement of Predictions
Bridges the logical development of the hypothesized relationships among the conceptual variable with the specific anticipated effects (foreshadowed earlier at the conceptual level) in terms of the concretely operationalized variables

Figure 8.1 *An inverted pyramid structure of the introduction.*

mentioning how the ways women and men interact can both reflect and reinforce gender disparities. It might then describe the main dependent variable of interest, noting that time spent speaking represents social dominance.

To establish the *significance of the research*, the first paragraph might also foreshadow how the interaction context can influence the status experienced by women and men in social and work-related interactions. This paragraph should conclude with a brief and clear statement of the major theses of the research and explain why this is novel and important. It is not sufficient to justify the work as something that has never been done before. As a colleague once commented to me, some things have not been done because they are not worth doing. Frame the work in a way that describes the value for advancing the field directly.

To create a solid empirical and theoretical foundation for the predictions, separate paragraphs should be devoted to the *specification of the "players"*. These paragraphs should highlight the literature that established the value and relevance of the major outcome of interest – speaking time – and, separately, on the effects of the independent variables – participant sex and the gendered nature of the social context – on social behavior. This part of the paper should reinforce the argument for the significance of the research that was initially noted in the introductory paragraph. The paragraphs should not be lists of previous findings; they should be smaller stories within your story. If you remove every parenthetical citation to specific studies, the text should be coherent with clear points and conclusions.

In a subsequent paragraph describing the *more complex dynamics* (see Figure 8.1), a logical argument about how and why the two independent variables can combine to affect speaking time should be developed. That paragraph should conclude with a general *conceptual* statement about how and why the gender-related nature of a context determines when a woman or a man will speak more in a mixed-sex interaction.

The next step is to provide readers with a brief *overview of the study procedure* that outlines what occurred in the research study. Readers will have a difficult time understanding the predictions if they do not know what you did in the study. Whereas the previous paragraphs discussed relevant literature and conceptual issues, this paragraph describes the ways these concepts were *operationalized* in the research.

Then, the conceptual analysis can be synthesized with the concrete manipulations and measures in another paragraph featuring the predictions. Unless the work is intended to be exploratory, there needs to be a clear *statement of predictions* that is supported by a compelling rationale. Stylistically, keep the introduction focused on developing a chain of logic that leads to the predictions. Avoid the temptation to discuss related and interesting issues that do not align directly with the logical sequence leading to the predictions. Simple is better than complex, and tangents should be avoided. If any sentence could begin with a phrase, "And another thing you might want to know is . . .", do not include it; keep readers focused on the main thread of logic. Bem (2000, p. 7) similarly recommends avoiding needless concepts and topics: "If a point seems tangential to your basic argument, leave it out. If you can't bring yourself to do this, put it in a footnote. Then, when you revise your manuscript, remove the footnote. In short, don't make your voice struggle to be heard above the ambient noise of cluttered writing . . . Write simply and directly".

The principle of simple and direct applies to sentence structure as well. Sternberg (1993) urges, "Write sentences that are readable, clear, and concise". Long and complex sentences are difficult for readers to parse; they are cognitively exhausting. One basic rule for effective writing is that no sentence deserves to be more than five lines long. It is short, declarative sentences that attract readers. Use them strategically to make key points.

To reduce the cognitive burden for readers and to enhance narrative flow, avoid long strings of citations in the text, as well as abbreviations and jargon, as much as possible. These unnecessarily interfere with the flow of the text. Try to limit in-text citations about a specific point to no more than three references – often a classic, a comprehensive, and a cutting-edge one. Readers are not looking to find an exhaustive list of references or to be impressed with how well you know the literature. They are reading the report to learn about your data and how it informs their own work. Also, avoid rhetorical questions because they put an unnecessary cognitive burden on the reader to come up with a transition that you could not.

Method

Different journals vary in the specific format they require for describing how the research was conducted. As noted earlier, it is in authors' best interests to conform completely to the requirements of the journal. Generally, though, the method(s) section of a report begins with a section explaining the participant sample – the number of participants, who the participants were (including relevant demographic information), where and how participants were recruited, and any other relevant information that could affect the interpretation of the results (e.g., participant attrition).

Some journals request separate measures and procedure sections; others allow the integration of measures within the narrative of the procedure. Either way, the basic information about the empirical precedent of a measure, the items that comprise a measure, response options, and data about the reliability and validity of a measure should be reported. Most effective procedure sections present what was done in a study chronologically and from the perspective of participants. What were participants told the study was about? What did they first do in the session, how was this presented to the participants, and why is this important to the study? When an element of the procedure involves the manipulation of an independent variable, explicitly tag this section with a subheading that alerts readers to the connection to the theoretical framework (e.g., "To manipulate the gender-related context of the interaction . . ."). Each of the specific conditions representing that independent variable should be explained.

Manipulation-check items, mediators, and dependent measures (appropriately tagged in terms of their relevance and role in the study) should be described at the point in which they were administered to participants in the study. To streamline the presentation of the material in this section and, often, to get under the word limit specified by the journal, more detailed information about the procedure can be provided in the supplementary materials. For published articles, supplementary materials are typically made available to readers online.

Critiques of practices in the social and behavioral sciences concerning how undisclosed flexibility in data collection and analysis produces misleading statistically significant results (false positives) in research reports have dramatically altered what is considered best practices. Simmons et al. (2010, 2011), who attracted broad attention for their insightful analysis of this issue, recommend including – when true – a 21-word statement in the method section: "We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study".

Results

The results section is not simply a list of statistical analyses and results. Instead, it is an integral component of the story you are telling. For complex results or a series of different types of findings, it is often useful to begin with an advance-organizer section that alerts readers to the sequence of material to be presented, explains what types of analyses were conducted (with what statistical packages), and reminds readers how they are relevant as manipulation checks or to the hypotheses developed in the introduction. There are two potential ways that the results could be structured. One way is to report a statistical effect (e.g., the overall difference [main effect] in speaking time for male and female participants) for each of the measures in the study and then for the next basic effect (e.g., the difference in the amount of time people talk in the masculine, feminine, and neutral topic conditions). The other way is to devote separate paragraphs to each dependent measure and report the relevant effects for just that variable (e.g., the main effects and interaction, from most general findings to most specific and complex). By convention, do the latter.

As recommended for other sections of the report, each paragraph in the results should begin with a topic sentence explaining what was done and why. The paragraph should end with a concluding sentence that summarizes the main point and facilitates the transition to the next paragraph. When an advance organizer is used in the results section, the topic and/or concluding sentence should orient readers to where they are on that roadmap. One way to assess the readability of the results section is to erase the statistical information to see if the narrative effectively conveys the meaning of the findings. Think of the statistical information as a kind of parenthetical information, like a reference, that mainly documents the validity of the statement that precedes it. Roediger (2007) agrees, stating that "[W]riters often lose their focus when reporting their results. The results section can be written using a format based on inferential statistics that makes for deadly dull reading . . .. A better strategy is for the author to make a story out of the descriptive statistics, telling what independent variables affected what dependent variables, and then provide F ratios (or other statistics) as supporting evidence that the effect cited in the prose is indeed significant".

Discussion

The discussion section has multiple purposes, including (a) summarizing the results and how they align with the hypotheses, (b) identifying and addressing any loose ends, and (c) suggesting promising directions for future research. It should also acknowledge limitations of the current research and offer concrete insights about how these limitations can be overcome. As with other sections, it should not be a list of these elements. It should be organized and told as an integrated story, one that dovetails with the story developed for the project as a whole.

The story that the discussion needs to tell must have a clear beginning, middle, and end. The beginning is what was done and what was found. The middle is the interpretation of the findings' novel contribution, including considering the potential value of unanticipated results and acknowledgement of limitations. The end involves offering concrete directions to extend the findings in significant ways and explaining the benefits of the work for theory development and application. The story needs to be structured (conforming to a predetermined outline), engaging, and coherent.

With respect to the first part of the discussion, as mentioned concerning the introduction, it is important to keep in mind that just because you found this topic important does not mean that others will find the project valuable. It is useful to open the discussion by articulating what the goals of the project were and emphasizing why they are important, particularly to the priorities of the target journal (e.g., theoretical advance and practical application in a specific area). When you get to the part that describes the results of your research, avoid simply re-hashing the findings from the results section. Keep the summary of the findings relatively brief, and foreground the most important findings, noting with brief subordinate clauses whether they support the predictions. You cannot assume that readers automatically appreciate the meaning of the findings. Because of the thought and effort you have invested in the project, you have a very close-up perspective that makes even small things loom large. To readers who are more distant from the project, it all looks relatively small. Therefore, you need to keep reminding readers, with details and crisp logic, how and why your work makes a valuable and unique contribution to the literature.

Often, authors' strong motivation to get their manuscript published tempts them to exaggerate some elements of the work while ignoring others. Although it is important to be a strong advocate for your research, it is also critical to be honest. Do not give attention only to supportive results while trying to hide findings that do not support predictions. Do not distort or misrepresent findings or overstate your conclusions. Besides being scientifically irresponsible, it is not in an author's best interest. Readers who detect such misrepresentation will tend to discount your findings. Earn and keep the reader's trust. Also, beyond readers' reactions to the specific study, overstating or misrepresenting results can have broader professional implications. You are in the profession for the long haul, and being recognized for honesty is critical in a profession that depends so much on the integrity of the researcher.

When addressing unanticipated results, there are two general approaches you might adopt. One is to try to explain away unsupportive findings by attributing them to specific aspects of the design, procedure, or measures used in the research. This argument is essentially that your hypotheses were right, but there were flaws in the execution that prevented you from being proven correct. The other approach is to assume that the data are right and that you were wrong. Both are acceptable, and people often adopt both perspectives in the discussion. However, it is helpful to transition as soon as possible to speculating about the meaning of these "loose ends". If readers pause and become distracted by thinking about their own alternative interpretations, it could unnecessarily cost you a gold coin. All research has loose ends, and many of these can suggest new directions for work that can move the field ahead. Thus, these unanticipated findings should be embraced by authors, who should try to decipher the clues they offer and propose to readers specifically how these loose ends can be productively pursued. By suspending the need to be right in this context, you can take advantage of the opportunity for expanding your perspective or discovering creative new insights.

Some journals require a subsection describing limitations of the work. This is not an invitation to list as many problems as you can. All studies have limitations. The objective here is to alert readers to the most important limitations, explain how and why they are limitations, and suggest concrete ways to address these limitations in future work. Phrases such as “more research is needed” are vacuous; no one expected this study to end the need for research. Make concrete suggestions to overcome limitations and address why it was a limitation. Some authors and journals prefer a separate future directions section, but guidance concerning the most productive future directions can also be woven into the consideration of unanticipated findings and limitations.

Common limitations to consider involve the nature of the sample (e.g., convenience samples of college students or people who opt in to online platforms). Even when representative samples are used, they are representative primarily of a particular population. Because research has longer traditions in some parts of the world, over 80% of findings are based on responses from WEIRD samples – samples from Western, educated, industrialized, rich, and democratic societies (Henrich et al., 2010). Yet, populations from WEIRD regions constitute only 12% of the world’s population. Because of the importance of recognizing such limitations to the generalizability of findings, Simons et al. (2017, p. 1123) have proposed that the “discussion section of all articles describing empirical research should include a statement of the *Constraints on Generality* (a ‘COG’ statement) that explicitly identifies and justifies the target populations for the reported findings”. However, including a statement acknowledging constraints on generalizability remain rare; currently, such a statement is not typically required by journals. Nevertheless, this proposition has stimulated considerable reflection in the field and, like other controversies that have emerged, may merit authors’ further attention. As noted earlier, publication expectations and standards do change.

Abstract

Some journals require structured formats with specified sections for the abstract (e.g., objective, methods, results, and conclusions); others request a single block of text. Also, journals vary substantially in their word limits. Although the abstract is located in the manuscript before the narrative, I have chosen to discuss this as the last part of this section for a reason – a primary objective of the abstract is to convey the essential information represented in each of the sections of the report (introduction, method, results, and discussion). Thus, a good way to begin is to “abstract” (in its meaning of to extract) key sentences in each section once they all have been written as an initial skeleton for the abstract. Of course, these sentences need to be reworded and synthesized for coherence.

A second objective of the abstract is to attract readership to your report. Electronic searching makes the inclusion of relevant key terms essential. However, once people find the abstract among a list of other relevant reports, readers will determine how valuable and interesting your work is. Therefore, it is important to include both a description of what is included in the report and why other scholars should read the paper. Even when the abstract has a very restrictive word limit, find a way to integrate a succinct but powerful statement of the theoretical and practical significance of the research into the abstract.

Revising the Report

Roediger (2007) notes, "Revision is the key to effective writing". Revisions typically need to be made at two main points. One is before the report is submitted to a journal; the other is after it has been reviewed. In the process of writing, seeking feedback about ideas, logical development, and style helps to fortify the material in the report as new material is added. For many of the reasons that lead people to experience writer's block, people are reluctant to seek feedback on their work. This feedback is critical because authors have difficulty taking the perspective of the reader. Also, writers are often resistant to hearing and benefitting from this feedback. I have been known to say, "My report is my baby, and it is beautiful to me no matter how ugly it appears to others". Still, I bristle at the thought that others see my baby as ugly. Beyond this initial reaction, authors are reluctant to see critiques of their work before they submit the report because being responsive to the feedback usually means restructuring arguments or eliminating sections of the text to which authors are attached. However, Bem (2000, p. 10) advises, "If your colleagues find something unclear, do not argue with them. Their suggestions for correcting the unclarities may be wrongheaded; but as unclarity detectors, readers are never wrong".

To facilitate interest in seeking comments from others and implementing revisions, authors should adopt an appropriate mindset. Two alternative approaches are performance and learning mindsets (Grant & Dweck, 2003). With a performance mindset, the focus is on the self; the main concern is on how well you are doing in terms of your own standards and others' impressions. With a learning mindset, the focus is on the activity and acquiring the information and the skills needed to master the task. A performance mindset interferes with writing, whereas a learning mindset facilitates it. Focusing on the importance of the material in the report for others is an effective way to anchor your writing in a learning mindset. Also, keep in mind that writing is a skill not simply a talent. There is no such thing as a "natural writer". Top writers have worked hard to master their craft, and everybody has the capacity to acquire the skills that will make them an effective writer. Adopting a learning mindset increases the likelihood that a person will seek input in the process of writing, deeply process the feedback, be responsive to comments in revising the report, produce a better report, and develop valuable life-long skills.

Revision also occurs, almost inevitably, because of the comments received when a report has undergone peer review. As mentioned earlier, the likelihood that a report will be accepted by a journal in its initially submitted form is virtually zero, regardless of the journal. Editors and reviewers almost always identify weaknesses and recommend changes. Authors' initial reactions are typically ones of frustration, anger, and reactance. These are not constructive reactions. As discussed in the earlier section about dealing with the emotions that contribute to writer's block, strategies that allow negative emotions to dissipate and that clear and open one's mind are valuable for gaining a proper perspective. Remember, writing is communication and, if readers do not understand something, it is the writer's responsibility to communicate more clearly. The reviewers and editor are not your enemies. They have devoted significant time, energy, and expertise to making your work better. Being open-minded allows authors to benefit from these recommendations.

Making revisions that are responsive to reviewers' comments is challenging because it involves at least three difficult steps. First, it requires overcoming initial affective responses to criticism that is often unexpected and seems overly harsh. For many of us, the response is a visceral one – more emotional than rational. Personally, after receiving reviewers' feedback, I cycle through several affective stages that parallel those in Kübler-Ross's (1969) classic model of grief: denial, anger, bargaining, depression, and acceptance. Second, when authors overcome their emotional reactance, they need to do the hard intellectual work of coming to understand specifically why something they worked hard on and believe is correct is perceived as problematic to others. Finally, once authors understand the reviewer's or editor's perspective, they need to provide new information or include additional explanation that successfully addresses the comment.

Although authors have the option of refuting a point made by the reviewer or editor, this should be done judiciously and generally after consulting with colleagues. Even when you might not fully agree with the point or question its importance, it is worth seriously considering a reviewer's or editor's request. If the reviewer or editor has a question about an aspect of the manuscript, other readers may have the same question. Effective writing recognizes and accommodates the perspective of the audience. In addition, in any response letter accompanying the revision of the manuscript, point-by-point explanations of how and where comments have been addressed in the text facilitates the re-review of the work. Typically, responding to these comments in a genuine way does make the work stronger.

Conclusion

The goal of this chapter was not simply to review the mechanics of writing a report, it was to create a broader understanding of the *process* of writing. *Doing* good research requires creativity and a range of methodological and statistical skills. Conducting research also has several intermittently rewarding features, such as selecting a question of personal interest, creating a design and procedure, and the discovery of answers through data analysis. *Writing* about the research you conducted in the report comes after you have learned the answers to the questions you asked. For many researchers, it is anticlimactic. For most, it is arduous. For all, it involves deferring gratification. While simply completing and submitting a manuscript is rewarding in many ways, those feelings pale compared to the joy of having your work accepted for publication. However, writing the report is an essential aspect of a sustained scholarly career. When you are done writing one report, it is time to begin another one.

Two main principles emphasized in this chapter are that writing is (a) a skill and (b) a form of communication. It is essential to embrace writing as a skill. Skills are developed through instruction, modeling, and practice. In her book about writing in political science, Baglione (2020) applies the metaphor of running a marathon to the process of writing a research report. Writing requires mental preparation, training, and the acquisition of skills and practice that maximize efficiency and effectiveness. A key objective in this chapter is to offer guidance in how to write a research report and to assist you in developing the kinds of habits and skills that will help you run your professional marathon.

Recognizing writing as a type of communication is another essential insight for becoming an effective writer. Adopting this perspective alerts you to the importance of being engaging, organized, logical, and careful in writing the report. Viewing scientific writing as a form of storytelling is valuable for recognizing the creative opportunities that exist. The story needs to be true, but non-fiction can be as stimulating as fiction. Viewing writing as a form of communication also makes you more open to feedback that is valuable for writing the piece at hand effectively and for developing general writing skills.

In conclusion, while we regularly think about *what* we write, I urge you to frequently pause to reflect on *why* we write. We write to convey a message about how our findings advance theory and can benefit society in practical ways. First-rate research described in a well-written report that is published in a high-profile outlet is the best way to achieve that. Advancement of our career is also an important consideration, but it is a by-product of the magnitude and originality of the contributions we make to our discipline and society. Writing is not primarily about you; it is about the message you bring. We all benefit when you do it well.

References

Amabile, T. M. (1983). Brilliant but cruel: Perceptions of negative evaluators. *Journal of Experimental Social Psychology*, *19*(2), 146–156. https://doi.org/10.1016/0022-1031(83)90034-3

Baglione, L. A. (2020). *Writing a Research Paper in Political Science: A Practical Guide to Inquiry, Structure, & Methods*, 4th ed. Sage.

Becker, H. S. (2008). Above all, write with clarity and precision. *Sociological Inquiry*, *78*(3), 412–416. https://doi.org/10.1111/j.1475-682X.2008.00247.x

Becker, H. S. (2020). *Writing for Social Scientists*, 3rd ed. University of Chicago Press.

Bem, D. J. (1987). Writing the empirical journal article. In M. P. Zanna & J. M. Darley (eds.), *The Compleat Academic* (pp. 171–201). Lawrence Erlbaum Associates.

Bem, D. J. (2000). Writing an empirical article. In R. J. Sternberg (ed.), *Guide to Publishing in Psychology Journals* (pp. 3–16). Cambridge University Press.

Bem, D. J. (2004). Writing the empirical journal article. In J. M. Darley, M. P. Zanna, & H. L. Roediger (eds.), *The Compleat Academic*, 2nd ed. (pp. 185–220). American Psychological Association.

Boice, R. (1993). Writing blocks and tacit knowledge. *Journal of Higher Education*, *64*(1), 19–54. https://doi.org/10.1080/00221546.1993.11778407

Byron, K., Khazanchi, S., & Nazarian, D. (2010). The relationship between stressors and creativity: A meta-analysis examining competing theoretical models. *Journal of Applied Psychology*, *95*(1), 201–212. https://doi.org/10.1037/a0017868

Chai, P. R., Carreiro, S., Carey, J. L., et al. (2019). Faculty member writing groups support productivity. *The Clinical Teacher*, *16*(6), 565–569. https://doi.org/10.1111/tct.12923

Dovidio, J. F. (2010). Publishing myths. *Dialogue*, *25*(1), 8–9.

Dovidio, J. F. & Gaertner, S. L. (2007). Communicating basic behavioral science beyond the discipline: Reflections from social psychology. In M. Welch-Ross & L. G. Fasig (eds.), *Handbook on Communicating and Disseminating Behavioral Science* (pp. 93–110). Sage.

Garcia, J. A., Rodriguez-Sánchez, R., & Fdez-Valdivia, J. (2020). Confirmatory bias in peer review. *Scientometrics*, *123*, 517–533. https://doi.org/10.1007/s11192-020-03357-0

Grant, H. & Dweck, C. S. (2003). Clarifying achievement goals and their impact. *Journal of Personality and Social Personal Psychology*, *85*(3), 541–553. https://doi.org/10.1037/0022-3514.85.3.541

Hawkins, S. A., Halpern, D. F., & Tan, S. J. (2007). Beyond university walls: Communicating and disseminating science outside the academy. In M. Welch-Ross & L. G. Fasig (eds.), *Handbook on Communicating and Disseminating Behavioral Science* (pp. 111–127). Sage.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Kim, K. & Johnson, M. K. (2015). Distinct neural networks support the mere ownership effect under different motivational contexts. *Social Neuroscience*, *10*(4), 376–390. http://dx.doi.org/10.1080/17470919.2014.999870

Kübler-Ross, E. (1969). *On Death and Dying*. Routledge. https://doi.org/10.4324/9780203010495

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology*, *64*(1), 2–17. https://doi.org/10.1002/asi.22784

Moore, D. A. & Schatz, D. (2017). The three faces of overconfidence. *Social and Personality Psychology Compass*, *11*, e12331. https://doi.org/10.1111/spc3.12331

Morewedge, C. K. & Giblin, C. E. (2015). Explanations of the endowment effect: An integrative review. *Trends in Cognitive Science*, *19*(6), 339–348. https://doi.org/10.1016/j.tics.2015.04.004

Nelson, L. D., Simmons, J., & Simonsohn, U. (2019). Psychology's renaissance. *Annual Review of Psychology*, *69*, 511–534. https://doi.org/10.1146/annurev-psych-122216-011836

Nosek, B. A., Beck, E. D., Campbell, L., et al. (2019). Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences*, *23*(10), 815–181. https://doi.org/10.1016/j.tics.2019.07.009

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Science*, *12*(6), 238–241. https://doi.org/10.1016/j.tics.2008.02.014

Petty, R. E. & Brinol, P. (2011). The elaboration likelihood model. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (eds.), *Handbook of Theories in Social Psychology*, *Volume 1* (pp. 224–245). Sage.

Rocco, T. S. & Hatcher, T. (2011). *The Handbook of Scholarly Writing and Publishing*. Jossey-Bass.

Roediger, H. L., III (2007). Twelve tips for authors. *APS Observer*. Available at: www.psychologicalscience.org/observer/twelve-tips-for-authors.

Simmons, J., Nelson, L., & Simonsohn, U. (2010). A 21-word solution. *Dialogue*, *26*(2), 4–7.

Simmons, J., Nelson, L., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128. https://doi.org/10.1177/1745691617708630

Sternberg, R. J. (1993). How to win acceptances by psychology journals: 21 tips for better writing. *APS Observer*. Available at: www.psychologicalscience.org/observer/how-to-win-acceptances-by-psychology-journals-21-tips-for-better-writing.

Sternberg, R. J. & Sternberg, K. (2010). *A Psychologist's Companion: A Guide to Writing Scientific Papers for Students and Researchers*, 5th ed. Cambridge University Press.

Wood, W. (2019). *Good Habits, Bad Habits: The Science of Making Positive Changes That Stick*. Picador.

PART II

The Building Blocks of a Study

9 Participant Recruitment

Jesse Chandler

Abstract

A strong participant recruitment plan is a major determinant of the success of human subjects research. The plan adopted by researchers will determine the kinds of inferences that follow from the collected data and how much it will cost to collect. Research studies with weak or non-existent recruitment plans risk recruiting too few participants or the wrong kind of participants to be able to answer the question that motivated them. This chapter outlines key considerations for researchers who are developing recruitment plans and provides suggestions for how to make recruiting more efficient.

Keywords: Data Collection, Experimental Design, Mail Surveys, Online Surveys, Sampling, Sample Size

Introduction

Deciding who to recruit into a study and how to recruit them is an important part of the research design process. The information gathered from a study feeds into some sort of decision such as which policy to enact, which product to sell, or which study to conduct next. A recruitment plan starts with a clear definition of the objectives of the study in mind. Researchers should consider five major factors in developing a study recruitment strategy:

1. **What population does the researcher want to understand?** Does the study need to include individuals with specific characteristics, and is it acceptable if some people who meet these criteria are excluded? The answer to this question determines the population of interest and what sample frame is used to represent them.
2. **What kind of inference is the researcher trying to make about the population?** Is the goal of the study to *describe* some aspect of the world or to develop a theory about how variables are related to each other? The answer to this question informs the sampling plan.
3. **How precise must the estimates obtained from the study sample be?** How wide can the confidence intervals around estimates be? Can estimates of effect sizes be biased so long as they are in the correct direction? The answer to the precision question informs both the sampling plan and the recruiting strategy.
4. **How quickly must data be collected?** The answer to this question mostly informs the recruiting strategy.

5. **What resources (e.g., time and money) are available for recruiting study participants?** The answer to this question will inform both the sampling plan and the recruiting strategy.

The answers to these questions will inform the sampling plan used to select potential participants, the strategies used to recruit them, and whether any special efforts to screen participants for eligibility or adapt recruitment strategies might be necessary. It is best practice to document study recruitment strategies and track how they evolve because they involve many interdependent decisions that have downstream consequences for analyzing and describing results.

Sometimes the design requirements uncovered through these questions will be in tension with each other. At a high level, there are obvious trade-offs between the first three considerations, reflecting aspects of the rigor or quality of the design, and the latter two, concerning speed and cost. As the maxim goes – you can have fast, good, or cheap; pick two. Assuming time and resource constraints, different kinds of rigor will create the need for further trade-offs. For example, if a researcher is interested in a less common population (such as African Americans or people with a specific job), it may be hard to find a sample that is large enough to provide precise estimates. It is possible that a researcher will have to either find a larger population or design a study that requires fewer participants to provide a meaningful result.

Defining the Population of Interest

An important first step in formulating a recruitment plan is developing a clear understanding of the population to be recruited. Four considerations should influence the decision of which population to study:

1. **The research question motivating the study.** Researchers can be interested in people in general or in a subpopulation defined by demographics, occupation, biographical experience, or any other characteristics.
2. **Sources of measurement error.** A study might use materials that are only appropriate for a specific subgroup even though the research question is more general. Survey measures might not be available in all languages or reading levels or may operationalize a study hypothesis with materials that are relevant only to a subset of the population.
3. **Simplicity.** A researcher may decide to deliberately exclude some people from participating in a study to simplify analysis. For example, brain imaging studies often restrict participation to right-handed individuals to avoid having to account for the different lateralization of brain function in left-handed participants.
4. **Sample access.** Researchers may realize that the population they are most interested in is too difficult to recruit and may need to modify their research question to match the populations that are available.

Creating a Sampling Plan

A sampling plan specifies how members of the target population will be selected for inclusion in the study. A sampling plan has three major properties: (1) the method by which potential participants are selected; (2) the frequency with which people of different types appear in the sample (i.e., the sample composition); and (3) the size of the sample.

How Will Participants Be Selected?

There are three broad sampling approaches that form a continuum of cost and rigor: a census, a probability sample, and a non-probability sample:

1. A census recruits the entire population and can be practical when the population of interest is small (e.g., employees at a firm, students at an institution, or people with a specific and unusual occupation – presidents of American universities).
2. A probability sample is any sample in which all potential survey participants can be identified and have a specified non-zero probability of being included in the study.
3. A non-probability sample is any sample that fails to meet the criteria of a probability sample, intentionally or not.

Probability samples are the gold standard in scientific research because they closely approximate the results of a census at a fraction of the cost. Perhaps more importantly, potential sources of bias can be corrected through statistical adjustment (for an overview, see Bethlehem, 2009). Uncertainty in measurements obtained from probability samples can also be quantified (e.g., through confidence intervals).

The simplest example of a probability sample (appropriately called a *simple random sample*) assigns everyone an equal probability of selection, as if sample members' names were written on papers drawn from a hat. More complex designs use a process called *stratification* (discussed in the section on sample composition) to ensure that the sample will have a specific composition. Other designs can account for multiple stages of selection, such as when clusters of people (e.g., schools, towns, or households) are sampled and then participants are sampled within them (see Daniel, 2011 for a taxonomy of probability sampling methods).

Non-probability samples are used by researchers who are unable to draw a probability sample. Probability samples require a comprehensive list of population members, called a *sampling frame*, which can be expensive or impossible to obtain, especially for populations defined by non-geographic traits, such as being diagnosed with a particular disease. In such cases, a researcher will either use an incomplete list of population members (leading those not on the list to have a zero probability of selection) or ask population members to self-identify as potential research participants (making it impossible to specify their probability of selection). Note, though, that sometimes samples can be simultaneously non-representative of the general population that a researcher cares about (such as people with a specific disease) and representative of an interesting subgroup (such as patients with that disease at a specific hospital).

Non-probability samples do not necessarily produce more biased results than probability samples, but whether they are biased and the degree to which they are biased is unknowable. This uncertainty may be tolerable under one of the following conditions:

- The proportion of the population covered by the non-probability sample is high and response rates are high, meaning there is less room for the sample to deviate from the population (Meng, 2018).
- The variables of interest in the study are uncorrelated with the probability of selection into the sample (Coppock, 2019).
- The research question focuses on associations between variables rather than point estimates. Associations between variables seem to be much less sensitive to population differences than point estimates (Pasek & Krosnick, 2010; Snowberg & Yariv, 2021).
- The study is testing a theory rather than describing the world or establishing generalizability to a population. The truth of a theory is established by its ability to predict what will happen under a specific set of conditions, but not by whether these conditions exist in the real world (Mook, 1983).

The decision to use a non-probability sample may also depend on how precise the answer to a research question must be. A researcher's concern about potential sample bias differs greatly when trying to discover if a relationship exists between two variables as opposed to when trying to establish the size of this relationship. This is especially true if the decision or action that follows from the data involve significant consequences for the data user or other stakeholders.

Deciding on Sample Composition

Instead of using a non-probability sample that recruits everyone who wants to participate (called a *convenience sample*), or a simple random sample that relies on chance to ensure that the sample resembles the population, researchers can take an active role in deciding the composition of their sample. Non-probability samples can deliberately select people who have the specific characteristics or recruit until quotas of people with different characteristics are met (called a *purposive sample*; Etikan et al., 2016). Similarly, probability samples can be *stratified*, meaning they are divided into subgroups based on characteristics or combinations of characteristics, with the desired number of participants drawn from each stratum.

The sample composition that researchers use should be determined by the research question:

- When estimating a population level effect, a researcher might select a sample that is representative of the population on characteristics that are likely to influence the size of the effect.

- When comparing subgroups of the population, a researcher might sample equal numbers of each group because comparing equally sized groups is more statistically efficient than comparing unequally sized groups (see Chapter 6 in this volume).
- When examining associations between continuous variables, a researcher might oversample participants with extreme scores on these variables. Adding participants with extreme scores increases the statistical power to detect associations between variables (Preacher et al., 2005; Sackett & Yang, 2000) and makes it easier to identify discrepant extreme cases that might have an undue impact on results.

A researcher may purposively sample specific people for any number of other legitimate reasons, particularly in qualitative research where the sample size is small. The researcher may select people that represent the "typical" member of a specific group, those with unusual experiences, or even those whose expertise in a topic makes their opinions especially informative. *Critical case sampling* is another useful technique in the formative stage of a research program; this approach targets participants whose responses can be used to make logical generalizations about the existence of an effect or it's boundary conditions (Patton, 2007). For example, in exploratory research, a researcher might seek out the participants who would be most likely to demonstrate that a phenomenon can *ever* happen before deciding whether conducting the study using a more representative sample is worth the effort.

Deciding on How Many People to Recruit

Researchers must estimate how many people to recruit into a study. A power analysis (see Chapter 5 in this volume) defines the minimum sample for the *analytic* data set and is the starting point for determining how many people to recruit. Additional participants will be needed because some will not respond (see the Maximizing Response Rates section later in this chapter for ways to limit non-response), drop out of the study, or be excluded during data cleaning (see Chapter 21 in this volume). The best estimates of how many people to recruit will come from direct experience with a specific population or method. Researchers without this experience can start with averages observed within their field and adjust their estimate based on how their study may differ.

To illustrate, a survey that requires an analytic sample of 100 participants probably requires an initial target of *at least* 112 complete responses to allow for participants excluded during data cleaning. About 2% of participants can be expected to skip any given question (Shoemaker et al., 2002). The quality of responses that participants provide will vary widely. An initial estimate that 10% (higher for web surveys) of responses will be unusable seems reasonable (for overviews, see Arthur et al., 2021; Curran, 2016) but could be higher or lower depending on participant motivation, the difficulty and sensitivity of the survey items, and the mode of the survey.

A survey that requires 112 responses will need to invite several hundred more people to participate because some people will not see the invitation or refuse to participate. For example, between 80% and 90% of people who begin an incentivized web survey can be expected to complete it (Liu & Wronski, 2018). Many invitees never begin a survey at all. The proportion of people invited to a survey who then complete it is low and might average somewhere between 40% and 50% (Anseel et al., 2010). Higher response rates can be expected for in-person (perhaps about 60%) and mail surveys (about 50%); expect lower rates for web (about 30%) and telephone surveys (about 20%; Lindemann, 2019). Survey response rates have been declining for years and all these estimates may be optimistic.

Estimates of response rates are bound to be imprecise and should be re-evaluated and adjusted once recruiting is underway. To accommodate these changes, a good sampling plan will draw an initial sample of potential participants and additional samples (sometimes called *replicates*) that can be used as needed. A smaller initial sample avoids recruiting too many people or recruiting only the easy-to-find people. The replicates provide an option to continue recruiting if it becomes clear that the analytic sample will be underpowered.

Developing a Recruiting Strategy

After the sampling approach has been selected, researchers need to decide how to recruit participants. A successful recruiting strategy will try to maximize survey response rates using the available time and resources. This process includes finding an adequate sample source, deciding on a general recruiting approach, determining how and how often to invite respondents to complete the study, and developing the contents of the study invitation.

List-Based Samples

Most probability samples and some non-probability samples begin with a list called a *sample frame*. The frame usually contains names of people but may enumerate some other unit (e.g., addresses of residences in a specific area and potentially active phone numbers) that the researcher will sample from instead. Sometimes it will include other characteristics of list members, such as contact information (e.g., phone numbers or email addresses) and demographic information. A researcher may need to obtain access to a list owned by a non-profit organization, firm, or other entity (for an overview of strategies for gaining such access, see Lindsay, 2005). Ideally the frame includes most or all the population of interest. The researcher should make sure they understand how the frame was generated and whether people are systematically excluded from it, a situation referred to as *coverage error*.

List-based samples offer several advantages over non-list-based samples. The contact information included in lists makes it much easier to recruit people to complete a study. Each sample member can be linked to their study materials, ensuring that only people who are eligible to complete the study can do so and that each person only participates once. Using a list also provides insight into how many people complete the study (i.e., the *response rate*). As discussed earlier, the characteristics of participants who do and do not complete the study can be compared for evidence of non-response bias, and the analytic sample can be weighted to match the characteristics of the sample frame. Probability samples require a list to satisfy the requirement that participants have a known and non-zero probability of selection.

For studies that use lists, simply locating people included on the list can be a challenge because contact information is constantly changing. About 15% of Americans move within any given year (Desilver, 2013) and, according to VerifyBee, an email validation service for email marketers, between 5% and 30% of email addresses, are invalid (VerifyBee, 2019). Researchers who use lists should be prepared to spend time cleaning and updating the information they contain. Depending on the sample type and resources available, locating could range from looking for participants on social media (LinkedIn, Facebook) or online directories, to a more involved approach, such as contacting likely relatives or sending field locators to known addresses (for a detailed description of one set of locating approaches, see Hall et al., 2003).

Non-List-Based Samples

Researchers who do not have access to a list of potential recruits can use other methods to recruit participants. Intercept samples (sometimes called *river samples*) recruit individuals from a flow of activity (e.g., people visiting a mall or a website). Intercept samples can efficiently target populations that are likely to congregate at a specific location. For example, parents could be recruited at children's museums or sporting events, and social media advertisements can be targeted at specific groups (Boas et al., 2020). People can be intercepted either in person or electronically by targeting electronic devices located within specified coordinates (called *geofencing*; see Haas et al., 2020). Website visitors can be intercepted through advertisements or pop-ups to complete studies (e.g., on social media). More recent iterations of river sampling will also intercept people using apps (e.g., Pollfish) and even those who dial an incorrect phone number (Reconnect Research; Levine et al., 2019).

Intercept sampling plans can be designed to randomly select recruits from a flow of activity, but they cannot produce a probability sample of any specific population because the sampling unit is the occurrence of an event, such as website visits or shopping experiences. For example, a store intercept study will over-represent frequent shoppers and exclude online-only shoppers. In some cases, this is perfectly fine because the research question concerns events (e.g., shopping experiences) and not individuals (e.g., consumers), but researchers should be careful presenting these results because the two are easily confused. A response rate cannot be calculated for intercept studies because the number of population members who have not been contacted is unknown. However, a cooperation rate can be calculated for the study (American Association for Public Opinion Research, 2016), which provides some information about how willing recruits were to participate in the study.

Volunteer samples can be recruited by posting requests for participants through print advertisements and flyers, posts on social media sites or discussion boards (e.g., Reddit; Shatz, 2017), email listservs, or even purpose-built websites (e.g., projectimplicit.net). The line between an intercept sample and a volunteer sample is sometimes blurry; volunteer sampling methods generally "pull in" people actively looking to participate in research, while intercept studies "push out" a call for participants to the general population (Antoun et al., 2016). Volunteer samples are less time-intensive to recruit and tend to attract more conscientious participants, probably because they attract people who are more motivated to participate in research (Antoun et al., 2016). The trade-off is that volunteers are highly self-selected based on their interest in research. A second consideration is that since the number of people who see the request for volunteers is unknown, the cooperation rate for a study can usually not be tracked when volunteer samples are used.

Network sampling, sometimes referred to as *snowball sampling* or *chain-referral sampling*, is a set of methods of recruiting research participants through referrals from previous participants. The network begins with an initial sample of the general population or an established list of people with the desired characteristics. As data are collected, each participant is asked to provide contact information of people they know who meet the study selection criteria. Usually, network sampling will produce a non-probability sample, though some designs can estimate the probability of selection and may closely approximate the results of a true probability sample (Heckathorn & Cameron, 2017). Network sampling is especially useful when recruiting from a population that is difficult to locate or that might be reluctant to identify themselves to a researcher before a peer can vouch for the legitimacy of the study.

Buying Access to Research Participants

If a researcher does not have access to their own participants, they can contract a research firm to provide them. Research firms often use complex sampling plans or even blend samples from different sources (Grenville & Berger, 2016). Researchers considering using a research firm should ask for their responses to the ESOMAR-36 (European Society for Opinion and Marketing Research, 2012), a set of questions designed to help standardize descriptions of sample provider practices and sample characteristics. These questions were created to help buyers evaluate different samples in a standardized and comprehensive way, and most reputable firms have prepared documents that will answer them.

Some firms offer access to probability samples of participants (for a recent list of well-established vendors, see Schonlau & Couper, 2017). Costs depend on the desired sample size and composition, but in the United States a 20-minute study using a probability sample of 1,000 to 2,000 participants could cost between $50,000 and $100,000. A portion of this cost is fixed and covers programming the survey, data cleaning, and some sort of weighting to adjust responses, while the remainder is variable and covers the costs of participant recruitment. One notable exception is the Time-Sharing Experiments for the Social Sciences (TESS; www.tessexperiments.org), a long-standing National Science Foundation-funded program that offers free access to a probability sample for survey experiments.

Other firms offer access to non-probability samples of participants, sometimes referred to as *online panels*. These firms maintain records that can help identify eligible participants and prescreen panelists for more specific criteria, if needed. Firms vary widely in the services they provide and their quality-control procedures. Higher-cost providers may handle survey programming and use complex screening criteria and statistical weights to make the final sample representative of the target population on some characteristics while lower-cost providers may not. To varying degrees, online panel providers manage their panel, retiring panelists after a set time, taking measures to prevent fraudulent or poor-quality responses, and recruiting new panelists to ensure a diverse sample. Again, costs will vary, but it is reasonable to expect that a non-probability sample, like the one described above, will cost between $10,000 and $50,000 to collect.

Some firms offer access to participants with few additional services, such as secure payment, rudimentary demographic screening, and a reputation system to weed out bad actors. Two examples of such services – Amazon Mechanical Turk and Prolific – are discussed in Volume 2 of this Handbook. These samples are usually used as convenience samples, but researchers can select participants who have specific characteristics from these samples. Recruiting the sample described above from one of these vendors likely costs under $10,000, though the researcher would have to handle survey programming, sample management, and data cleaning.

Combining Recruiting Strategies

Researchers can combine samples together to increase sample sizes or to compensate for defects unique to each sample. Non-probability samples can easily be combined with each other simply by entering responses into the same data set. Combining very different types of non-probability samples can be useful because results that replicate across different sample sources provide at least some evidence that findings are generalizable. Researchers can also use statistical techniques to combine probability samples with non-probability samples, maintaining the lack of bias of the probability sample while also enjoying the efficiency of non-probability recruitment methods (Gellar et al., 2020). To illustrate, a survey of homeless people based on a roster of shelter residents will not cover those who do not use shelters. To overcome this limitation, a researcher could augment the roster with an intercept sample or network sample of homeless people recruited from streets or encampments (Dennis, 1991).

Maximizing Response Rates

One of the primary goals of any recruiting plan is to maximize response rates. When a sample frame is small, a high response rate may be the only way to attain the analytic sample size needed for a study. A high response rate also usually reduces non-response bias of survey results, though this is not always the case (Groves & Peytcheva, 2008).

Another goal is to maximize the efficiency of data collection. Cost per complete is one measure of efficiency that can be calculated by summing the fixed cost of creating survey materials and total variable costs of labor, incentives, and other direct costs for attaining each completed response and dividing this total by the total number of responses. Estimating the cost per complete can help researchers evaluate how best to allocate limited resources, even though estimates of how different design options might affect response rates are bound to be imprecise (for an example see Williams et al., 2018).

When developing a recruitment strategy, researchers should consider four general factors that can influence the response rates and cost per complete:

1. How will participants be made *aware* that a study is available for them to complete? Contact information might be missing or inaccurate. Even if the message is delivered, it may go unread – approximately 80% of people say they screen their calls and ignore unfamiliar numbers (McClean, 2020), 20% ignore mail that looks like advertising (Mazzone & Pickett, 2011), and 75% of emails are never opened (MacDonald, 2021).
2. How can the *expected benefits* of completing the survey be maximized? Benefits can mean incentives (e.g., money) or other less tangible rewards (e.g., feeling good about helping others, complying with social norms, or fulfilling a moral obligation; Bosnjak et al., 2005). Expected benefits can differ from actual benefits if people believe that they are unlikely to be realized (Dillman, 1978).
3. How can the perceived *costs* of completing the survey be minimized? Costs can include participant time and hard-to-quantify hassle factors such as complicated instructions or questions about sensitive topics that discourage participation.
4. How do sample *demographics* affect likely response rates? Some populations are harder to locate or are more reluctant than others to participate in research. Populations can also differ in the subjective value they place on different costs and benefits, leading them to respond differently to recruitment strategies (Groves et al., 2000).

With these considerations in mind, there are several design choices that influence response rates. Where available, this account reports the average impact of the design choices observed in meta-analytic reviews along with the amount of evidence upon which these estimates are based. Results are reported for mail and web-based studies because mail is the most well-studied mode and web studies are most frequently used by researchers.

Designing a Study That People Will Participate in

As study design decisions can influence response rates, it is worth thinking about response rates while designing the study. In particular, shorter studies consistently produce higher response rates than the longer ones (Edwards et al., 2009; Göritz, 2014; Liu & Wronski, 2018; Mavletova & Couper, 2015). The effect of survey length is non-linear (Reyes, 2020), but it is likely that 20 minutes is at the outside of what most people will tolerate (Revilla, 2017; Revilla & Höhne, 2020).

People are less willing to complete surveys that feel difficult or uncomfortable. Including open-ended questions reduces response rates to mail surveys by about 70% (Edwards et al., 2009) and reduces response rates to mail surveys by about 12% (Liu & Wronski, 2018). Including sensitive questions reduces response rates by about 10% (Edwards et al., 2009). Conversely, "interesting" questionnaires had response rates that were twice as high as "uninteresting" questionnaires (Edwards et al., 2009; see also Marcus et al., 2007). One common recommendation is to start a study with easier or more enjoyable tasks to warm up respondents. Supporting this claim, beginning with an open-ended question leads to response rates that are 5% lower than using a multiple-choice question (Liu & Wronski, 2018).

Selecting Appropriate Study Modes for the Target Population

Though web surveys are nearly ubiquitous, researchers can also collect data through other modes, including in-person interview, mail, telephone, or unconventional electronic modes (e.g., text messages or apps; De Bruijne & Wijnant, 2014). Aside from the strengths and limitations of each mode for data collection (for an overview, see Tourangeau, 2018), the mode also influences response rates. Web surveys have response rates that are about 11% lower than what could be attained through non-electronic means of data collection (Daikeler et al., 2020), even among youth and young adults (Cantrell et al., 2018).

Web surveys can also produce biased results. Though web penetration increases each year, web users remain younger, more educated, higher earners than non-web users (Schumacher & Kent, 2020), and some groups still face serious barriers to getting online. The psychological characteristics of web users also differ from those of non-users (Marcus & Schutz, 2005; Rogelberg et al., 2003), even when studying populations with high levels of Internet penetration (Callegaro et al., 2014).

Data can be collected in more than one mode to improve study response rates and sample diversity (Messer & Dillman, 2011). Different survey modes increase the odds of reaching people with partially incorrect or missing contact information. They may also appeal to people who might be willing to complete a survey in one mode but not in another. Different modes should be offered sequentially because offering a choice of response modes up front reduces response rates (Medway & Fulton, 2012). For efficiency, multimode studies usually begin with a low-cost mode (e.g., a web survey) and move to more expensive modes (e.g., mail or phone surveys) for sample members who have not responded (for a detailed overview of designing mixed-mode surveys, see Dillman & Edwards, 2016).

Sending a Prenotification Letter

One common practice is to send a prenotification message (usually a letter) that alerts recipients about the survey, explains its purpose and contents, and addresses any concerns about security or privacy. Prenotifications increase response rates to mail surveys by about 50% compared to sending a study invitation with no prenotification (Edwards et al., 2009).

The mode used to send a prenotification (and initial study invitation) can and often should differ from the mode the study is offered in. A mixture of mail, phone, and email notifications ensures that a researcher can reach potential participants even if some or all their contact information is missing and can push them to a web study. Mail can be an especially powerful recruiting mode because a letter signals the legitimacy of a study (Dillman, 2017). A meta-analysis found that sending advance letters prior to a telephone survey increased response rates by an average of 8 percentage points (de Leeuw et al., 2007) and other studies have observed similar benefits for web surveys (Lawes et al., 2021; Sakshaug et al., 2019).

Giving People Many Opportunities to Respond

Sending reminders to complete the study is among the most effective ways to increase response rates, though with diminishing returns for each additional reminder sent (see Sánchez-Fernández et al., 2012; but for an exception, see Van Mol, 2017). On average, sending reminders increases response rates to mail surveys by about 35% (Edwards et al., 2009). For web surveys, reminders increase response rates by about 30% (Göritz, 2014). Interestingly, for both mail and electronic surveys, the value of sending a prenotification seems to exceed that of sending one or more reminder messages (for a direct comparison that supports this observation, see Andreadis et al., 2020).

A recruiting plan should specify how often people will be reminded about the study and when. The time and day of the week have small and inconsistent effects on overall response rates to web surveys, but early in the work week seems best (see Griggs et al., 2021). Phone surveys should try to contact potential recruits on different times and days and be mindful of how participant availability may differ across time zones. Study reminders should also be spaced out to avoid annoying potential participants. About half of the people who respond to web surveys do so within the first day, with almost everyone else replying within a week (Reynolds et al., 2009; Sauermann & Roach, 2013).

Offering Incentives

The easiest way to motivate people to complete a study is to offer them an incentive, and this is by far the most studied predictor of response rates. Offering a monetary incentive increased response rates in mail surveys by 87% (Edwards et al., 2009), but had a smaller effect for web surveys, increasing response rates by 20% (Göritz, 2006).

Larger monetary incentives usually have larger impacts on response rates. However, increasing payment yields diminishing returns. Paying people a dollar (as opposed to nothing) increases response rates by about 5 percentage points, but each additional dollar above $5 increases response rates less than 2 percentage points; each increase above $10 increases response rates by about 1 percentage point (across 55 trials; Mercer et al., 2015 replicated by Jia et al., 2021).

Non-monetary incentives (including merchandise, entrance into lotteries, and sharing the results of the study) are less effective than monetary incentives, increasing response rates in mail surveys by only 15%, and response rates are unaffected by the value of the non-monetary incentives (Edwards et al., 2009). Offering to share the results of the survey has essentially no effect on response rates for mail surveys (Edwards et al., 2009). Non-monetary incentives also have little to no effect in web surveys (Daikeler et al., 2020).

Lotteries are especially popular with researchers because they are inexpensive, but they rarely improve response rates (Singer & Ye, 2013). Offering a few very large prizes may be more effective than offering many smaller prizes (Conn et al., 2019). The effect of lottery incentives is also larger in electronic surveys when respondents are immediately told if they have won, perhaps because the expected value of the incentive is higher if they do not have to worry that a reward message will be overlooked or blocked as spam (Tuten et al., 2004).

For mail and phone surveys, incentives have the largest impact on response rates when they are prepaid unconditionally to respondents, increasing response rates by 61% relative to paying the same amount upon completion of the study (Edwards et al., 2009; see also Mercer et al., 2015). One reason for this finding may be that a prepaid incentive creates a social obligation to reciprocate by completing the survey. Prepaid incentives have less impact in web surveys (Coopersmith et al., 2016; Edwards et al., 2009), perhaps because of reduced trust or social obligation online. A major disadvantage of prepaid incentives is that they are costly because they are paid to everyone rather than only those who complete the study.

Designing an Effective Survey Invitation

Survey invitations are persuasive communications. As such, the psychological principles that underpin persuasive writing are important (Groves et al., 1992). Since participating in a study takes time and effort, a successful study invitation requires strong arguments for participating. The invitation and reminders could include the following information:

- the purpose and contents of the study, framed in a way to highlight questions that are personally relevant or interesting to respondents (Marcus et al., 2007)
- any incentives for completing the study
- the length of the study, especially if it is short (Trouteaud, 2004)
- a deadline for when the study should be completed (Porter & Whitcomb, 2003).

Claims about the intangible benefits of completing the study for the researcher, participant, or society have little effect on response rates (Edwards et al., 2009).

Some research has found that more peripheral cues can make survey invitations more persuasive. These cues have small effects:

- the institution responsible for the data collection, especially if it is credible and trustworthy (Edwards et al., 2009)

- normative information that others have already responded (Porter & Whitcomb, 2003)
- an appeal for help (Petrovčič et al., 2016; Trouteaud, 2004)
- humor, if appropriate for the topic and audience (Rath et al., 2017).

If people will be contacted more than once, the contents of each message should be different. Changing the wording of each reminder seems to improve response rates by about 30% compared to using the same language (Sauermann & Roach, 2013). As one example, the original contact can provide a detailed explanation of who is conducting the study, why it is being conducted, and how the recipient was selected. Later messages can be shorter and emphasize different reasons why the recipient should respond.

It is important to signal the importance of the potential respondent and the researcher's willingness to invest resources in securing a response. Personalization is one way to communicate importance. Addressing potential respondents by name improves response rates by about 10% for mail surveys and 25% for electronic surveys (Edwards et al., 2009). Handwritten signatures on invitation letters and handwritten addresses had somewhat larger effects, boosting response rates by more than 25% for mail surveys (Edwards et al., 2009). Similarly, more costly methods of sending study invitations produce higher response rates. Mail outperforms email, but sending mail surveys by special delivery (e.g., certified mail) increases response rates by an additional 50%, and including a stamped (as opposed to a business reply) return envelope increases response rates by 25% (Edwards et al., 2009).

For emailed survey invitations and forum posts, the subject line is an opportunity to make a first impression. It will often determine whether the survey invitation is opened and read or just deleted. People report disliking posts with subject lines that are uninformative or that seem like "clickbait" and prefer straightforward subject lines that emphasize the purpose of the study (Brenner et al., 2020). Embedding the first question within the survey invitation itself also seems to increase the proportion of recipients who start and complete the study (Liu & Inchausti, 2017)

Adjusting Expectations Based on Sample Demographics

In general, men, younger people, people with low or extremely high incomes, people with less education, and single people are all less likely to respond to surveys (Reyes, 2020), though this can vary with study mode (for an overview, see Goyder 2019). These differences can inform estimates of the likely response rates that could be obtained from a population or be used to focus recruiting efforts on people who are less likely to respond.

Researchers should consider whether their sample has any special challenges and design their recruitment strategy with these challenges in mind. Some populations, such as homeless youth, are highly mobile and difficult to remain in contact with (Eyrich-Garg & Moss, 2017), suggesting that studies of this population should focus on raising awareness of the study. Other populations, such as physicians, may value monetary incentives less and require researchers to either offer very large incentives or other reasons to complete the study.

Screening Participants for Study Eligibility

When a study requires participants to have specific characteristics, the eligibility of potential recruits must be confirmed through a screening process. Screening should be conducted with care because there is ample evidence that survey fraud can be a problem. In a stark illustration, a study of medical research participants recruited using newspapers and Craigslist (an American classified advertisements website) found that 14% of participants admitted to fabricating a health condition to gain eligibility to a paid clinical trial (Devine et al., 2013). Importantly, the prevalence of fraudulent responses in a data set will depend on the ratio of fraudulent participants to truly eligible sample members. When recruiting for studies with participants who have uncommon characteristics, it is easy to end up with a sample where the majority of responses are fraudulent (for a detailed discussion, see Chandler & Paolacci, 2017).

One way to reduce the impact of survey fraud is to increase the number of truly eligible participants by recruiting only people who are likely to have the desired characteristics. Ideally, a researcher would have a list of people who meet the study criteria and screen them only to verify their eligibility. When such a list is unavailable, researchers can find other ways to focus recruiting efforts on people who are likely eligible for the study. For example, a survey of people receiving government benefits could focus recruiting efforts on people with lower incomes, accepting the loss of people who have recently experienced a reduction in income as a good tradeoff for excluding many people who are certainly ineligible.

Screening instruments can be designed to minimize participant fraud. When possible, eligibility criteria should not be disclosed to recruits before eligibility is measured. Ideally, recruits will not even be aware that they were screened at all. For example, people could be recruited into a short study that includes the screening questions, with eligible participants immediately routed to a second survey if they qualify (Springer et al., 2016).

Testing and Adapting the Recruiting Plan

Testing is an important part of developing a recruiting plan. Large-scale studies are often preceded by a dress rehearsal in which a small sample is selected, recruited, and administered the study to identify any major operational problems. Smaller studies can benefit from at least pilot testing the full recruiting procedure in the same way that they might pilot test the research instrument. Testing can uncover ambiguity in recruiting protocols, misunderstandings between team members, or other problems before the study is launched. Testing may also reveal that assumptions about the cost, response rate, or effort involved in conducting the study are inaccurate.

Once the study has begun, researchers should keep a close eye on important metrics such as the overall response or completion rate; the degree to which the completed sample represents the underlying population; and the cost per complete response (overall and perhaps for key subgroups). Based on these metrics, researchers can adjust the recruitment strategy to focus on specific modes or channels, change the content of survey invitations, or target certain participants with additional contact attempts or larger incentives. For example, when trying to minimize the cost, researchers can focus their efforts on the groups that are more likely to respond; when trying to improve representativeness, they can focus on those least likely to respond (Groves & Heerenga, 2006; for a detailed treatment see Schouten et al., 2020).

Documenting and Reporting Recruitment Methods

At a minimum, a research report should specify if a sampling frame was used, how participants were selected, the steps used to recruit them, and any available statistics about response rates or cooperation rates that might inform whether non-response bias is a concern. The method used to determine the numerator and denominator of the response rate should also be specified because there are many plausible methods of calculating response rates (for definitions and formulas, see American Association for Public Opinion Research, 2016).

If a sampling frame is not used, researchers should try to identify the target population, factors such as the time and location of recruiting efforts that may influence sample composition, and any constraints on the generality of their findings that may result from these decisions (Simons et al., 2017). They should also report any exclusion criteria that were used to screen participants.

Researchers could also document recruitment plans and materials (such as through a preregistration plan), the effort expended to recruit, and results. Preserving records of what was done and what did and did not work prevents institutional knowledge from being lost when staff leave the research group. Online repositories and the increasing willingness of journals to publish supplementary materials also make it easy to share detailed recruiting methods with others. Omitting these details can make it difficult for other researchers to understand the study population, directly replicate a finding, or improve their own knowledge of how to successfully recruit research participants.

Conclusion

The recruiting plan is a critical part of the research design, determining the kinds of research questions that a study can answer, whether the design is precise enough to answer them, and how expensive and time-consuming it will be to collect data.

A study without a well-conceived recruiting plan can be ruined in ways that are only discovered during data analysis or peer review. Poorly planned studies can also end up consuming time, money, and other resources that could have been devoted to other projects.

There is an old aphorism that plans are useless, but planning is indispensable. Plans are not useless, but it is true that the act of planning itself offers the most value. The ease with which a recruitment strategy can adapt to changing circumstances will depend on how well planned the strategy was. Planning forces researchers to set priorities and to identify important goals, define milestones, and find obstacles in the recruiting process. Planning also provides a sense of how different design decisions affect each other, and the trade-offs between cost, quality, and speed that each decision might entail. These details are important when working within a budget, a research schedule, a power calculation, and other constraints, but they can be easy to overlook when under the time pressures caused by actively collecting data.

Acknowledgments

Thanks to Rachel Berry, Brian Brutzman, Laetitia Lemonie, Jeewon Oh, and the editors of this volume for valuable comments on the first draft of this chapter.

References

American Association for Public Opinion Research (2016). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*, 9th ed. American Association for Public Opinion Research.

Andreadis, I. (2020). Text message (SMS) pre-notifications, invitations and reminders for web surveys. *Survey Methods: Insights from the Field, Special Issue: Advancements in Online and Mobile Survey Methods*. https://doi.org/10.11587/DX8NNE

Anseel, F., Lievens, F., Schollaert, E., & Choragwicka, B. (2010). Response rates in organizational science, 1995–2008: A meta-analytic review and guidelines for survey researchers. *Journal of Business and Psychology*, *25*(3), 335–349. https://doi.org/10.1007/s10869-010-9157-6

Antoun, C., Zhang, C., Conrad, F. G., & Schober, M. F. (2016). Comparisons of online recruitment strategies for convenience samples: Craigslist, Google AdWords, Facebook, and Amazon Mechanical Turk. *Field Methods*, *28*(3), 231–246. https://doi.org/10.1177/1525822X15603149

Arthur, W., Jr., Hagen, E., & George, F., Jr. (2021). The lazy or dishonest respondent: Detection and prevention. *Annual Review of Organizational Psychology and Organizational Behavior*, *8*, 105–137. https://doi.org/10.1146/annurev-orgpsych-012420-055324

Bethlehem, J. (2009). *Applied Survey Methods. A Statistical Perspective*. John Wiley & Sons.

Boas, T. C., Christenson, D. P., & Glick, D. M. (2020). Recruiting large online samples in the United States and India: Facebook, Mechanical Turk, and Qualtrics. *Political*

Science Research and Methods, *8*(2), 232–250. https://doi.org/10.1017/psrm.2018.28

Bosnjak, M., Tuten, T. L., & Wittmann, W. W. (2005). Unit (non) response in web-based access panel surveys: An extended planned-behavior approach. *Psychology & Marketing*, *22*(6), 489–505. https://doi.org/10.1002/mar.20070

Brenner, P. S., Cosenza, C., & Fowler, F. J., Jr. (2020). Which subject lines and messages improve response to e-mail invitations to web surveys? *Field Methods*, *32*(4), 365–382. https://doi.org/10.1177/1525822X20929647

Callegaro, M., Villar, A., Yeager, D., & Krosnick, J. A. (2014). A critical review of studies investigating the quality of data obtained with online panels based on probability and nonprobability samples. In Callegaro, M., Baker, R., Bethlehem, J., et al. (eds.), *Online Panel Research: A Data Quality Perspective* (pp. 23–53). John Wiley & Sons.

Cantrell, J., Bennett, M., Thomas, R. K., et al. (2018). It's getting late: Improving completion rates in a hard-to-reach sample. *Survey Practice*, *11*(2). https://doi.org/10.29115/SP-2018-0019

Chandler, J. J. & Paolacci, G. (2017). Lie for a dime: When most prescreening responses are honest but most study participants are impostors. *Social Psychological and Personality Science*, *8*(5), 500–508. https://doi.org/10.1177/1948550617698203

Conn, K. M., Mo, C. H., & Sellers, L. M. (2019). When less is more in boosting survey response rates. *Social Science Quarterly*, *100*(4), 1445–1458. https://doi.org/10.1111/ssqu.12625

Coopersmith, J., Vogel, L. K., Bruursema, T., & Feeney, K. (2016). Effects of incentive amount and type of web survey response rates. *Survey Practice*, *9*(1), 1–10.

Coppock, A. (2019). Generalizing from survey experiments conducted on Mechanical Turk: A replication approach. *Political Science Research and Methods*, *7*(3), 613–628. https://doi.org/10.1017/psrm.2018.10

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

Daikeler, J., Bošnjak, M., & Lozar Manfreda, K. (2020). Web versus other survey modes: An updated and extended meta-analysis comparing response rates. *Journal of Survey Statistics and Methodology*, *8*(3), 513–539. https://doi.org/10.1093/jssam/smz008

Daniel, J. (2011). *Sampling Essentials: Practical Guidelines for Making Sampling Choices*. SAGE Publications.

De Bruijne, M. & Wijnant, A. (2014). Improving response rates and questionnaire design for mobile web surveys. *Public Opinion Quarterly*, *78*(4), 951–962. https://doi.org/10.1093/poq/nfu046

de Leeuw, E. D., Callegaro, M., Hox, J., Korendijk, E., & Lensvelt-Mulders, G. (2007). The influence of advance letters on response in telephone surveys. *Public Opinion Quarterly*, *71*(3), 413–443. https://doi.org/10.1093/poq/nfm014

Dennis, M. L. (1991). Changing the conventional rules: Surveying homeless people in non-conventional locations. *Housing Policy Debate*, *2*(3), 699–732. https://doi.org/10.1080/10511482.1991.9521070

Desilver, D. (2013). Chart of the week: Americans on the move. *Pew Research Center*, November 22. Available at: www.pewresearch.org/fact-tank/2013/11/22/chart-of-the-week-americans-on-the-move/.

Devine, E. G., Waters, M. E., Putnam, M., et al. (2013). Concealment and fabrication by experienced research subjects. *Clinical Trials*, *10*, 935–948. https://doi.org/10.1177/1740774513492917

Dillman, D. A. (1978). *Mail and Telephone Surveys: The Total Design Method*. John Wiley & Sons.

Dillman, D. & Edwards, M. (2016). Designing a mixed mode survey. In C. Wolf, D. Joye, T. Smith, & Y.-C. Fu (eds.), *SAGE Handbook of Survey Methodology* (pp. 255–268). SAGE Publications.

Dillman, D. (2017). The promise and challenge of pushing respondents to the Web in mixed-mode surveys. *Survey Methodology*, Statistics Canada, Catalogue No. 12-001-X, Vol. 43, No. 1. Available at: www.statcan.gc.ca/pub/12-001-x/2017001/article/14836-eng.htm.

Edwards P. J., Roberts I., & Clarke M. J., et al. (2009). Methods to increase response to postal and electronic questionnaires. *Cochrane Database of Systematic Reviews*, MR000008. https://doi.org/10.1002/14651858.MR000008.pub4

Etikan, I., Musa, S. A., & Alkassim, R. S. (2016). Comparison of convenience sampling and purposive sampling. *American Journal of Theoretical and Applied Statistics*, *5*(1), 1–4. https://doi.org/ 10.11648/j.ajtas.20160501.11

European Society for Opinion and Marketing Research (2012). 28 Questions to help buyers of online samples. Available at: https://swiss-insights.ch/wp-content/uploads/2020/05/ESOMAR-28-Questions-to-Help-Buyers-of-Online-Samples-September-2012.pdf.

Eyrich-Garg, K. M. & Moss, S. L. (2017). How feasible is multiple time point web-based data collection with individuals experiencing street homelessness? *Journal of Urban Health*, *94*(1), 64–74. https://doi.org/10.1007/s11524-016-0109-y

Gellar, J., Hughes, S., Delannoy, C., et al. (2020). *Multilevel Regression with Poststratification for the Analysis of SMS Survey Data (No. c71d456bbf9f4026988e1a8107df4764)*. Mathematica Policy Research.

Göritz, A. S. (2006). Incentives in web studies: Methodological issues and a review. *International Journal of Internet Science*, *1*(1), 58–70.

Göritz, A. S. (2014). Determinants of the starting rate and the completion rate in online panel studies. *Online Panel Research: Data Quality Perspective, A*, 154–170. https://doi.org/10.1002/9781118763520.ch7

Goyder, J. (2019). *The Silent Minority: Non-Respondents in Sample Surveys*. Routledge.

Grenville, A. & Berger, R. (2016). *Rivers, Routers and Reality.* Maru/Blue. Available at: https://static1.squarespace.com/static/5a5d2933a8b2b0f49d244724/t/5d3b65ac163abf00017f5e0f/1564173741975/Rivers%2C+Routers%2C+and+Reliability+-+A+Test+of+Sample+Sources.+Data+Quality%2C+and+Reliability+-+Maru+Blue+Whitepaper.pdf.

Griggs, A. K., Smith, A. C., Berzofsky, M. E., et al. (2021). Examining the impact of a survey's email timing on response latency, mobile response rates, and breakoff rates. *Field Methods*, March 30. https://doi.org/10.1177/1525822X21999160

Groves, R. M., Cialdini, R. B., & Couper, M. P. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly*, *56*(4), 475–495. https://doi.org/10.1086/269338

Groves, R.M. & Heeringa, S. G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A*, *169*(3): 439–457. https://doi.org/10.1111/j.1467-985X.2006.00423.x

Groves, R. M. & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: a meta-analysis. *Public Opinion Quarterly*, *72*(2), 167–189. https://doi.org/10.1093/poq/nfn011

Groves, R. M., Singer, E., & Corning, A. (2000). Leverage–saliency theory of survey participation: Description and an illustration. *Public Opinion Quarterly*, *64*(3), 299–308. https://www.jstor.org/stable/3078721

Haas, G. C., Trappmann, M., Keusch, F., Bähr, S., & Kreuter, F. (2020). Using geofences to collect survey data: Lessons learned from the IAB-SMART study. *Survey Methods: Insights from the Field*, December 10. https://doi.org/10.13094/SMIF-2020-00023

Hall, E. A., Zuniga, R., Cartier, J., et al. (2003). *Staying in Touch: A Fieldwork Manual of Tracking Procedures for Locating Substance Abusers in Follow-Up Studies*, 2nd ed. UCLA Integrated Substance Abuse Programs

Heckathorn, D. D. & Cameron, C. J. (2017). Network sampling: From snowball and multiplicity to respondent-driven sampling. *Annual Review of Sociology*, *43*, 101–119. https://doi.org/10.1146/annurev-soc-060116-053556

Jia, P., Furuya-Kanamori, L., Qin, Z. S., Jia, P. Y., & Xu, C. (2021). Association between response rates and monetary incentives in sample study: A systematic review and meta-analysis. *Postgraduate Medical Journal*, *97*(1150), 501–510. https://dx.doi.org/10.1136/postgradmedj-2020-137868

Lawes, M., Hetschko, C., Sakshaug, J. W., & Grießemer, S. (2021). Contact modes and participation in app-based smartphone surveys: Evidence from a large-scale experiment. *Social Science Computer Review*, March 11. https://doi.org/10.1177/0894439321993832

Levine, B., Krotki, K., & Lavrakas, P. J. (2019). Redirected inbound call sampling (RICS) telephone surveying via a new survey sampling paradigm. *Public Opinion Quarterly*, *83*(2), 386–411. https://doi.org/10.1093/poq/nfz024

Lindeman, N. (2019) What is the average survey response rate? Available at: https://surveyanyplace.com/average-survey-response-rate/.

Lindsay, J. (2005). Getting the numbers: The unacknowledged work in recruiting for survey research. *Field Methods*, 17(1), 119–128. https://doi.org/10.1177/1525822X04271028

Liu, M. & Inchausti, N. (2017). Improving survey response rates: The effect of embedded questions in web survey email Invitations. *Survey Practice*, *10*(1), 1–6. https://doi.org/10.29115/SP-2017-0005

Liu, M. & Wronski, L. (2018). Examining completion rates in web surveys via over 25,000 real-world surveys. *Social Science Computer Review*, *36*(1), 116–124. https://doi.org/10.1177/0894439317695581

MacDonald, S. (2021). The science behind email open rates (and how to get more people to read your emails). Available at: www.superoffice.com/blog/email-open-rates/.

Marcus, B. & Schütz, A. (2005). Who are the people reluctant to participate in research? Personality correlates of four different types of nonresponse as inferred from self- and observer ratings. *Journal of Personality*, *73*(4), 959–984. https://doi.org/10.1111/j.1467-6494.2005.00335.x

Marcus, B., Bosnjak, M., Lindner, S., Pilischenko, S., & Schütz, A. (2007). Compensating for low topic interest and long surveys: a field experiment on nonresponse in web surveys. *Social Science Computer Review*, *25*(3), 372–383. https://doi.org/10.1177/0894439307297606

Mavletova, A. & Couper, M. P. (2015). A meta-analysis of breakoff rates in mobile web surveys. In A. Mavletova & M. P. Couper (eds.), *Mobile Research Methods:*

Opportunities and Challenges of Mobile Research Methodologies (pp. 81–98). Available at: www.jstor.org/stable/j.ctv3t5r9n.11.

Mazzone, J. & Pickett, J. (2011). *The Household Diary Study: Mail Use & Attitudes in FY 2010*. The United States Postal Service.

McClean, C. (2020) Most Americans don't answer cellphone calls from unknown numbers. *Pew Research Center*, December 14. Available at: www.pewresearch.org/fact-tank/2020/12/14/most-americans-dont-answer-cellphone-calls-from-unknown-numbers/.

Medway, R. L. & Fulton, J. (2012). When more gets you less: A meta-analysis of the effect of concurrent web options on mail survey response rates. *Public Opinion Quarterly*, *76* (4), 733–746. https://doi.org/10.1093/poq/nfs047

Meng, X. L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Annals of Applied Statistics*, *12*(2), 685–726. https://10.1214/18-AOAS1161SF

Mercer, A., Caporaso, A., Cantor, D., & Townsend, R. (2015). How much gets you how much? Monetary incentives and response rates in household surveys. *Public Opinion Quarterly*, *79*, 105–129. https://doi.org/10.1093/poq/nfu059

Messer, B. L. & Dillman, D. A. (2011). Surveying the general public over the Internet using address-based sampling and mail contact procedures. *Public Opinion Quarterly*, *75*, 429–457. https://doi.org/10.1093/poq/nfr021.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist*, *38* (4), 379–387. https://doi.org/10.1037/0003-066X.38.4.379

Pasek, J. & Krosnick, J. A. (2010). Measuring intent to participate and participation in the 2010 census and their correlates and trends: Comparisons of RDD telephone and non-probability sample Internet survey data. *Statistical Research Division of the US Census Bureau*, 15, 2010.

Patton, M. Q. (2007). Sampling, qualitative (purposive). *The Blackwell Encyclopedia of Sociology*. John Wiley & Sons.

Petrovčič, A., Petrovčič, G., & Manfreda K. L. (2016). The effect of email invitation elements on response rate in a web survey within an online community. *Computers in Human Behavior*, *56*, 320–329. https://doi.org/10.1016/j.chb.2015.11.025

Porter S. R. & Whitcomb M. E. (2003). The impact of contact type on web survey response rates. *Public Opinion Quarterly*, *67*, 579–588.

Preacher, K. J., Rucker, D. D., MacCallum, R. C., & Nicewander, W. A. (2005). Use of the extreme groups approach: a critical reexamination and new recommendations. *Psychological Methods*, *10*(2), 178–192. https://doi.org/10.1037/1082-989X.10.2.178

Rath, J. M., Williams, V. F., Villanti, A. C., et al. (2017). Boosting online response rates among nonresponders: a dose of funny. *Social Science Computer Review*, *35*(5), 619–632. https://doi.org/10.1177/0894439316656151

Revilla, M. (2017). Analyzing survey characteristics, participation, and evaluation across 186 surveys in an online opt-in panel in Spain. *Methods, Data, Analyses*, *11*(2), 135–162. https://doi.org/10.12758/mda.2017.02

Revilla, M. & Höhne, J. K. (2020). How long do respondents think online surveys should be? New evidence from two online panels in Germany. *International Journal of Market Research*, *62*(5), 538–545. https://doi.org/10.1177/1470785320943049

Reyes, G. (2020). Understanding nonresponse rates: Insights from 600,000 opinion surveys. *The World Bank Economic Review*, *34* (Supplement), S98–S102. https://doi.org/10.1093/wber/lhz040

Reynolds, S., Sharp, A., & Anderson, K. (2009). Online surveys: Response timeliness and issues of design. Available at: www.researchgate.net/profile/Anne-Sharp/publication/266471625_Online_Surveys_Response_timeliness_issues_of_design_Online_Surveys_Response_timeliness_issues_of_design/links/556ba5e708aec22683037c00/Online-Surveys-Response-timeliness-issues-of-design-Online-Surveys-Response-timeliness-issues-of-design.pdf.

Rogelberg, S. G., Conway, J. M., Sederburg, M. E., et al. (2003). Profiling active and passive nonrespondents to an organizational survey. *Journal of Applied Psychology*, *88*(6), 1104. https://doi.org/10.1037/0021-9010.88.6.1104

Sackett, P. R. & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, *85*(1), 112–118. https://doi.org/10.1037/0021-9010.85.1.112

Sakshaug, J. W., Cernat, A., & Raghunathan, T. E. (2019). Do sequential mixed-mode surveys decrease nonresponse bias, measurement error bias, and total bias? An experimental study. *Journal of Survey Statistics and Methodology*, *7*(4), 545–571. https://doi.org/10.1093/jssam/smy024

Sánchez-Fernández, J., Muñoz-Leiva, F., & Montoro-Ríos, F. J. (2012). Improving retention rate and response quality in web-based surveys. *Computers in Human Behavior*, *28* (2), 507–514. https://doi.org/10.1016/j.chb.2011.10.023

Sauermann, H. & Roach, M. (2013). Increasing web survey response rates in innovation research: An experimental study of static and dynamic contact design features. *Research Policy*, *42*(1), 273–286. https://doi.org/10.1016/j.respol.2012.05.003

Schonlau, M. & Couper, M. P. (2017). Options for conducting web surveys. *Statistical Science*, *32*(2), 279–292. https://10.1214/16-STS597

Schouten, B., Peytchev, A., & Wagner, J. (2020). *Adaptive Survey Design*. Chapman and Hall/CRC Press.

Schumacher, S. & Kent, N. (2020). 8 charts on internet use around the world as countries grapple with COVID-19. *Pew Research Center*, April 2. Availabale at: www.pewresearch.org/fact-tank/2020/04/02/8-charts-on-internet-use-around-the-world-as-countries-grapple-with-covid-19/.

Shatz, I. (2017). Fast, free, and targeted: Reddit as a source for recruiting participants online. *Social Science Computer Review*, */35*(4), 537–549. https://doi.org/10.1177/0894439316650163

Shoemaker, P. J., Eichholz, M., & Skewes, E. A. (2002). Item nonresponse: Distinguishing between don't know and refuse. *International Journal of Public Opinion Research*, *14*(2), 193–201. https://doi.org/10.1093/ijpor/14.2.193

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128. https://doi.org/10.1177/1745691617708630

Singer, E. & Ye, C. (2013). The use and effects of incentives in surveys. *The ANNALS of the American Academy of Political and Social Science*, *645*(1), 112–141. https://doi.org/10.1177/0002716212458082

Snowberg, E. & Yariv, L. (2021). Testing the waters: Behavior across participant pools. *American Economic Review*, *111*(2), 687–719. https://10.1257/aer.20181065

Springer, V., Martini, P., Lindsey, S., & Vezich, I. (2016). Practice based considerations for using multi-stage survey design to reach special populations on Amazon's Mechanical Turk. *Survey Practice*, *9*(5), 1–8. https://doi.org/10.29115/SP-2016-0029

Tourangeau R. (2018). Choosing a mode of survey data collection. In D. Vannette & J. Krosnick (eds.), *The Palgrave Handbook of Survey Research*. Palgrave Macmillan. https://doi.org/10.1007/978-3-319-54395-6_7

Trouteaud, A. R. (2004). How you ask counts: A test of internet-related components of response rates to a web-based survey. *Social Science Computer Review*, *22*, 385–392. https://doi.org/10.1177/0894439304265650

Tuten T.L., Galesic M., & Bosnjak, M. (2004). Effects of immediate versus delayed notification of prize draw results on response behavior in web surveys: An experiment. *Social Science Computer Review*, *22*, 377–384. https://doi.org/10.1177/0894439304265640

Van Mol, C. (2017). Improving web survey efficiency: the impact of an extra reminder and reminder content on web survey response. *International Journal of Social Research Methodology*, *20* (4), 317–327. https://doi.org/10.1080//13645579.2016.1185255

VerifyBee (2019). How to fix an invalid email address. *VerifyBee*, June 10. Available at: https://verifybee.com/how-to-fix-an-invalid-email-address.

Williams, D., Edwards, S., Giambo, P., & Kena, G. (2018). Cost effective mail survey design. In Proceedings of the Federal Committee on Statistical Methodology Research and Policy Conference, Washington, DC, December.

10 Informed Consent to Research[†]

David S. Festinger, Karen L. Dugosh, Hannah R. Callahan, and Rachel A. Hough

Abstract

This chapter focuses on informed consent, the cornerstone of conducting ethical human subjects research. It presents a brief history of the origins of informed consent to research and reviews codes, guidelines, and regulations that have been established in response to ethical violations carried out in the name of science. The chapter reviews the essential elements of consent (i.e., intelligence, knowingness, and voluntariness) and discusses challenges that researchers may encounter within each of these areas. Importantly, it approaches consent as an ongoing process rather than a one-time-event and presents practical and empirically supported strategies that researchers can apply to assess and enhance individuals' capacity, understanding, and autonomy as it pertains to research participation. Additional topics discussed include assent to research that involves children, electronic and multimedia consent, and consent to research using biospecimens.

Keywords: Informed Consent, Capacity, Understanding, Autonomy

Introduction

A great number of medical and behavioral advancements, including the development of treatments for deadly diseases like malaria, syphilis, and hepatitis, entailed years of research and testing with human subjects. Unfortunately, many were attained at the expense of marginalized and highly vulnerable populations such as asylum inmates, prisoners, people with intellectual disabilities, and non-institutionalized racial and ethnic minorities (Layman, 2009). Individuals were frequently involved in clinical trials without ever being informed of their involvement in research and, as a result, were unaware of what was happening to them. In short, accompanying its positive scientific contributions and societal benefits, the history of human subjects research retains a legacy of abuse in which the rights of individuals were subjugated to the goal of scientific progress. This chapter provides a discussion of informed consent, the critical role it plays in research, and practical strategies that researchers can employ to ensure that consent to research is truly informed.

[†] This chapter is dedicated to the memory of Dr. David Festinger who devoted much of his career to advancing research ethics in marginalized populations, particularly individuals who have substance use disorders.

https://doi.org/10.1017/9781009010054.011 Published online by Cambridge University Press

Historical Developments in Human Subjects Research

Prior to the middle of the twentieth century, there was essentially no regulatory oversight of human subjects research (Layman, 2009). Examples of research atrocities that occurred in the name of science include the Nazi medical experiments (Shuster, 1997), Tuskegee Syphilis Study (Jones, 1993), Milgram's Obedience and Individual Responsibility Study (Milgram, 1974), and the Human Radiation Experiments (Faden, 1996), all of which sharpened public awareness of the potential for abuse of or harm to research participants. In response to these and other ethical infractions, the United States and other nations have adopted and continue to revise regulatory policies to protect human research participants.

The first major international document to provide guidelines on research ethics, the Nuremberg Code (International Military Tribunal, 1950), stipulates 10 principles for conducting ethical research. A central tenet of the Code is voluntary consent in which individuals must:

- have the capacity to consent to participation
- not be coerced to participate
- be informed of and understand the research's purpose and procedures and the risks and benefits of the research.

The Nuremberg Code also mandates that researchers minimize suffering, ensure that risks do not significantly outweigh potential benefits, use appropriate study designs, and uphold participants' freedom to withdraw from the research at any time.

Strongly rooted in the Nuremberg Code, the World Medical Association adopted the Declaration of Helsinki in 1964. The Association has revised the Declaration periodically since its original formulation with the last update in 2013 (World Medical Association, 2013). The Declaration outlines a number of ethical principles for human subjects medical research. It more clearly delineates the appropriate ethical conditions for medical research and more closely recalls the moral obligations of physicians to their patients under the Hippocratic Oath. For instance, it states that research should be scientifically grounded, and the benefits should be proportionate to the risks.

The US Congress subsequently passed the National Research Act of 1974, largely responding to public outcry over the infamous Tuskegee Syphilis Study (Layman, 2009). The Act created the country's first federal body to oversee bioethics in research, the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. Two significant outcomes of the Commission were the requirements that researchers obtain informed consent and oversight from an institutional review board (IRB).

The Belmont Report (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979) focuses on three principles that underlie the ethical conduct of research:

- "Respect for persons" recognizes the autonomy and dignity of individuals and the need to protect those with diminished autonomy, such as children and individuals with cognitive impairments.

- “Beneficence” refers to the obligation to protect persons from harm by maximizing benefits and minimizing risks they experience.
- “Justice” entails the fair distribution of the benefits and burdens of research to all.

In 1991, the US Department of Health and Human Services (HHS) and 15 other federal departments and agencies issued the Federal Policy for the Protection of Human Subjects, generally referred to as the “Common Rule” (Electronic Code of Federal Regulations, 2018). The Common Rule provides a comprehensive regulatory framework for HHS-conducted or -supported research involving human subjects. The Common Rule specifies provisions for human subjects’ research protections, researchers, IRBs, and sponsoring institutions. In 2000, the HHS established the Office for Human Research Protections to regulate and provide oversight for research captured under the Common Rule. Importantly, the Common Rule was revised in 2019 to address the momentous changes in research that had occurred since 1991 (e.g., internet and biospecimens).

Role of Informed Consent

Considering milestones in human subjects’ protections from the Nuremberg Code up through the revised Common Rule, informed consent has emerged as the keystone of research involving human subjects. Informed consent requires that the following constructs have been met:

- Intelligence: the individual has the cognitive capacity to make a rational and informed decision regarding their participation in the research based on the study-related information presented to them.
- Knowingness: the individual fully understands the information presented to them and the implications of participation on their well-being.
- Voluntariness: the individual decides, free from coercion and undue influence, whether to participate in the research.

Informed consent is a process that occurs in the context of a researcher–participant relationship characterized by respect and candor. It involves ensuring that the participant understands the information that is presented to them and that researchers welcome and address questions that the participant may have, disclose new information that may arise in the course of the research, and re-obtain consent, if appropriate (Gupta, 2013).

Components of Informed Consent

Mandatory Elements of Informed Consent

The Common Rule delineates both the general requirements for the consent process and basic elements that must be included or omitted during the consent process (Electronic Code of Federal Regulations, 2018). The following six general requirements and nine basic elements are mandatory for all informed consent procedures and documents:

General Requirements:

(1) An individual or their legally authorized representative must provide legally effective informed consent before they can participate in research.

(2) An individual or their representative must be given sufficient opportunity to discuss and consider whether to participate in research, and consent is obtained under conditions that minimize coercion or undue influence.

(3) Consent and study information must be presented in language understandable to the individual or their representative.

(4) An individual or their representative must be given information that a reasonable person would consider necessary to make an informed decision on whether to participate in the research and sufficient opportunity to discuss that information.

(5) Consent forms must begin with a concise and focused presentation of key information that is most likely to aid an individual or their representative in understanding why one might or might not choose to participate in the research. This information must be provided in sufficient detail and must be presented in a way that optimizes understanding.

(6) The consent process and form cannot include exculpatory language that asks or gives the impression of asking the individual or their representative to give up their legal rights.

Basic Elements:

(1) A statement that the study involves research and a description of the study procedures, purpose, duration of the individual's participation, and any experimental procedures involved.

(2) Identification of any reasonably foreseeable physical, psychological, and social risks and discomforts to the individual

(3) Identification of any reasonably foreseeable benefits to the individual or others.

(4) Disclosure of alternative treatments that could be beneficial to the individual.

(5) An explanation of the extent to which confidentiality of records will be maintained and who will have access to the individual's information.

(6) An explanation of any compensation and treatment for research-related injuries resulting from studies that pose more than minimal risk to individuals.

(7) Information on whom to contact for questions about the research, the individual's rights as a participant, and research-related injuries or complaints.

(8) Statements that participation is voluntary and refusal to participate or discontinue participation will not involve penalties or loss of benefits to which the individual is otherwise entitled.

(9) If the research involves the collection of identifiable private information or identifiable biospecimens, a statement disclosing whether or not this information may be used or distributed for future research studies.

The general requirements and basic elements are intended to help prospective participants or their legally authorized representatives understand why they may or may not want to participate in research. They are designed to provide additional assurance to researchers, IRBs, and sponsoring institutions that all participants have received sufficient information about the study and that their decision to participate was intelligent, knowing, and voluntary. Importantly, additional provisions are required for conducting research with specific vulnerable populations, including pregnant women, fetuses, and neonates (Subpart B); prisoners (Subpart C); and children (Subpart D; Electronic Code of Federal Regulations, 2018).

If the research meets certain conditions, an investigator may seek a waiver or alteration of informed consent from the IRB conducting the review. Research that presents no more than minimal risk to participants, where "minimal risk" refers to risk that one could encounter in daily life or during routine physical or psychological examinations, may be granted such a provision provided that the rights and well-being of the participants are upheld. An IRB may remove the requirement to obtain signatures from participants, omit or modify some or all the mandatory elements, or completely eliminate the consent process altogether.

An investigator may also obtain broad consent (rather than study-specific informed consent) for research involving the storage, maintenance, and/or use of identifiable private information or biospecimens for secondary research (Maloy & Bass, 2020). With broad consent, researchers obtain participants' permission to use their information or biospecimens in studies that may be conducted in the future. Broad consent requires the same basic elements as standard informed consent but does not require certain information that is unknown at the time (e.g., study benefits and procedures).

Informed Consent for Children: Assent

The Common Rule provides special protections to children involved in research (Electronic Code of Federal Regulations, 2018). Children are defined as persons who have not reached the legal age for consent to research. This generally refers to individuals under the age of 18, but state or local laws may mandate different definitions. As the decision-making capacity of children may not yet be fully developed, they may be susceptible to undue influence and coercion, and as such they cannot provide informed consent to participate in research. In these cases, permission to participate in research from one or both biological or adoptive parents or legal guardian(s) is required.

In addition to obtaining consent from the legal guardian, researchers must obtain assent from the child when they can provide assent. Their capacity to assent is judged by evaluation of age, maturity, and cognitive capacity. Assent is defined as a child's explicit, affirmative agreement to participate in the research. The IRB determines the circumstances under which assent is solicited, obtained, and documented. Furthermore, the IRB may waive assent if the child is judged to lack capacity to assent or if the research holds the possibility of direct benefit to the well-being of the child, such as in a clinical trial.

Although the Common Rule is specific to research in the United States, other countries have adopted similar guidelines and regulations regarding the requirement of assent. However, there is a great degree of heterogeneity across countries (e.g., Lepola et al., 2016). For this reason, it is important for researchers to be aware of the guidelines and regulations related to assent in the country in which the research is being conducted.

Electronic Informed Consent

More recent technological advancements have led to the adoption of new procedures for conducting research, including the use of electronic informed consent (US Food and Drug Administration, 2016). Electronic informed consent (eIC) refers to an informed consent procedure conducted using electronic devices and/or internet technology. The use of eIC may help to simplify the documentation and storage of consent, promote the inclusion of participants from varied geographic regions (and who speak different languages), facilitate the re-administration of informed consent, and enhance comprehension of the presented information through the use of visual aids and advanced graphics (De Sutter et al., 2020).

In response to these new practices, the Food and Drug Administration (FDA) and the Office for Human Research Protections (OHRP; a division of the HHS) released guidance on the use of eIC in 2016 (US Food and Drug Administration, 2016). The guidelines indicate that eIC must adhere to the same regulatory requirements as in-person informed consent. Potential participants' questions related to the consent form should be answered through a phone call, electronic message, or video conference. Electronic signatures can serve in the place of written signatures or oral consent and should be digitally documented and stored. For remote eIC, the onus of obtaining informed consent is placed on the researcher. The FDA and OHRP suggest employing a protocol for identity confirmation, such as verification of an official identification document or correctly answering personalized questions.

Despite its many potential advantages, eIC raises a number of ethical concerns. Although eIC may increase access to research participation for some, it may preclude participation for those who are unaccustomed to or who lack the required technology as well as for individuals with visual or motor impairments (US Food and Drug Administration, 2016; De Sutter et al., 2020). The use of eIC may also introduce additional risks related to falsification of identity, data privacy, and research crossing jurisdictional boundaries. Furthermore, the electronic nature and lack of human connection may diminish the researcher–participant relationship. These complex issues require further review and discussion by regulatory agencies, IRBs, and researchers.

Ensuring Consent Is Informed

Empirical research on informed consent has demonstrated its many limitations in adequately ensuring that participants' decisions to participate in research are truly intelligent, knowing, and voluntary.

Intelligence

The first aspect of informed consent is intelligence. In this context, intelligence (i.e., "capacity to consent") refers to the ability to understand the information presented during the informed consent process and use it to make a decision about participation

(National Institutes of Health, 2009). This definition of intelligence relates to both ethical standards and legal requirements in research and clinical care and is invoked to promote safeguards for populations with impaired capacity to consent (i.e., who may be particularly vulnerable to coercion or undue influence; Appelbaum & Grisso, 2007). Capacity to consent can be adversely affected by numerous conditions including psychiatric, neurological, metabolic, and substance-use disorders as well as medications, trauma, and infections. It is important to understand that capacity exists on a spectrum and is changeable.

Over the course of the twentieth century, revelations of research abuses involving cognitively impaired individuals generated debate regarding the appropriateness of including them in research (Carlson, 2013). In one such incident, children with intellectual disabilities at the Willowbrook State School on Staten Island in New York City were purposefully infected with hepatitis between the 1950s and 1970s for the purpose of advancing treatment for the virus. Some have considered this group to be too vulnerable and the procurement of their informed consent prohibitively complicated. However, the Belmont principle of justice may be violated by both over-inclusion and under-inclusion of certain groups in research regardless of vulnerability, as the risks and benefits of research should be equally distributed across all groups. Excluding the participation of individuals with cognitive disabilities constitutes discrimination, withholds potential benefits to the individual (i.e., through clinical trials), and undermines scientific advancements that may serve the group (National Institutes of Health, 2009). Though more consideration is necessary, research specifically targeted to and including vulnerable populations should be undertaken when ethically appropriate, such as when there is direct possibility of benefit (Forster & Borasky, 2018).

For research involving individuals with impaired capacity to consent, the Common Rule mandates that IRBs take care to ensure that the selection of research participants is equitable and that extra safeguards uphold their rights and protect their well-being. However, federal regulations do not mandate how to assess capacity to consent or circumstances in which a legally authorized representative should be appointed. In this absence, research and medical communities have contributed their expertise. Consultation with experts, whether during IRB review, in community advisory boards, or otherwise, is crucial to ensure that the highest ethical standards are upheld in the informed consent process (HHS et al., 2016).

Assessment of Capacity to Consent

Specific methods of determining capacity and specific thresholds representing sufficient capacity must be established by both the researcher and the IRB. For research that may involve individuals with diminished capacity to consent, an established protocol is necessary, and IRBs may require documentation of capacity. Thresholds for capacity may vary depending on the type of research being conducted. For instance, studies that are deemed to be of minimal risk may have lower thresholds than high-risk studies (Forster & Borasky, 2018). This comports with standards that are used in clinical care (Appelbaum & Grisso, 2007) and the

FDA's regulations governing the inclusion of children in research. Lastly, it is important to continuously reassess participants' capacity as it may fluctuate throughout the course of a study.

Researchers may conduct an informal screen at the beginning of the informed consent process. This can involve conversations with the individual prior to the disclosure of study information (Appelbaum & Grisso, 2007; National Institutes of Health, 2009). They may also administer a questionnaire at the conclusion of the consent process to assess understanding of important information, such as the purpose of the research, potential risks and benefits, and the duration of participation. A formal assessment of capacity may be warranted whenever the researcher questions a participant's capacity to provide informed consent.

The MacArthur Competence Assessment Tool for Clinical Research (MacCAT-CR; Appelbaum & Grisso, 2001) is the most widely used formal assessment of capacity to participate in research. It is a well-validated 15- to 30-minute, semi-structured interview customizable to different research protocols. In a review of 12 different instruments, Dunn et al. (2006) found that the MacCAT-CR was the only instrument that adequately assessed the four different elements of decisional capacity (see Box 10.1). It is reliable and valid in a wide range of populations with diminished capacity, including individuals with schizophrenia, Alzheimer's disease, and diabetes (Palmer et al., 2005). It also has several limitations, including its length, training requirements for administration and interpretation, and a lack of standardization relating to its customization of items (Gilbert et al., 2017).

Another widely used instrument to assess capacity to consent is the University of California Brief Assessment of Capacity to Consent (UBACC; Jeste et al., 2007). This brief, 10-item tool is validated for use in a range of populations including individuals with schizophrenia (Jeste et al., 2007), substance-use disorder (Martel et al., 2018), and Alzheimer's disease (Seaman et al., 2015). The UBACC is relatively easy to incorporate into the consent process as it is quick to administer and does not require lengthy training. It is customizable like the MacCAT-CR, but its assessment of capacity may be viewed as less comprehensive. Researchers and IRBs

Box 10.1 The four elements of capacity to consent (Appelbaum & Grisso, 2007)

The criteria upon which capacity to consent to research is assessed are:

(1) **Understanding of the information**: A potential participant demonstrates a reasonable ability to comprehend and retain the information presented during the informed consent process.

(2) **Appreciation of the situation**: The individual applies the presented information to their own circumstances and understands the possible implications of it upon them.

(3) **Logical reasoning**: The individual utilizes the presented information to deliberate the options and reach a logical decision.

(4) **Communication of choice**: The individual definitively and consistently indicates their decision through verbal or non-verbal means.

must consider the advantages and disadvantages of available assessment instruments when deciding which to use. Factors including study population, the degree of risk, and the setting and circumstances of recruitment can help to inform this decision.

A recurring concern in both informal and formal assessment of capacity to consent is the lack of standardization in determining thresholds for capacity. In informal screens, what different researchers consider the demonstration of sufficient capacity may vary significantly. In relation to formal assessment, there is no official consensus on when it is appropriate to administer an instrument, which instrument(s) to employ, how to interpret the results, or what constitutes sufficient and insufficient capacity. In the absence of federal guidance, researchers should consult with experts in the study population, including IRBs, community advisory boards, and other researchers.

Legally Authorized Representatives

Some individuals and groups, such as those with severe intellectual disabilities, will lack the appropriate degree of capacity to provide informed consent. As mentioned above, blanket exclusion from participation in research would violate the Belmont principle of justice. Thus, another Belmont principle, that of respect for persons, is invoked in a two-fold manner – though individuals should have the utmost autonomy over research-related decisions, in cases in which the individual lacks capacity, it is equally important that they are not excluded (Appelbaum & Grisso, 2007). For this reason, it may be necessary to involve legally authorized representatives (LARs) to make decisions about research participation on behalf of individuals who have limited capacity.

The regulations and ethical principles applicable to informed consent obtained from a LAR are identical to those applicable to the potential participant. The Common Rule, however, does not mandate who may serve in the role of LAR and instead defers to state and local laws as well as institutional policies. Where there are no laws, institutional policies, or legally binding agreements (e.g., power of attorney), the role of the LAR should be filled by the individual who would serve to make decisions on behalf of a patient lacking capacity to consent to clinical care in the institution where the research is taking place. Many states and institutions have policies mandating a hierarchy of such potential surrogates that generally give preference to close relatives (e.g., California Legislative Information, 2003).

Facilitating Autonomy for Individuals with Diminished Capacity

There are steps that researchers can take to enhance the autonomy of individuals with cognitive deficits to enable them to provide informed consent for themselves (Evans et al., 2020; National Institutes of Health, 2009). This may apply to individuals with degenerative conditions (e.g., dementia), fluctuating conditions (e.g., schizophrenia), or relatively mild impairments in cognition (e.g., mild intellectual disability) but not individuals who consistently demonstrate diminished capacity to consent. The researcher and the IRB must exercise expert judgment in determining the extent

https://doi.org/10.1017/9781009010054.011 Published online by Cambridge University Press

to which enhancing an individual's capacity to consent is appropriate and ethically sound.

Modifications to the informed consent process can augment a potential participant's capacity to consent by compensating for cognitive shortcomings (Evans et al., 2020). For instance, one study found that a multimedia consent process improved scores on both the MacCAT-CR and the UBACC for individuals with schizophrenia (Jeste et al., 2009). For individuals with intellectual disability, altering language and communication techniques as well as involving close relations may improve their ability to consent (Ho et al., 2017). Presenting information more gradually and appropriating more time for decision-making can also enhance an individual's ability to process information and form a decision (National Institutes of Health, 2009).

For individuals with degenerative or fluctuating capacity, researchers should attempt to obtain informed consent when impairment is minimized, such as during the initial stages of a condition or when psychiatric symptoms decline (Appelbaum & Grisso, 2007; National Institutes of Health, 2009). For these individuals, it may also be appropriate for researchers to seek process consent, whereby consent is re-obtained after the initial consent process, as necessary (Evans et al., 2020). Process consent gives participants or their LARs the option to re-evaluate the former's participation in the study, especially if their medical status has changed. For example, researchers conducting a clinical trial involving end-of-life cancer patients sought informed consent from a participant's LAR when the patient was deemed to have lost capacity (Davies et al., 2018).

In conclusion, conducting research with vulnerable individuals who exhibit diminished capacity to consent may raise ethical concerns and prove more timely and complex than research with the general population. Nevertheless, it is critical to include such groups and individuals to ensure that they too reap the benefits of scientific research and advancements.

Knowingness

Participants face a growing list of risks and benefits to research participation as research protocols become more complex and involved (Morán-Sánchez et al., 2016; Sonne et al., 2013). Knowingness, referring to an individual's comprehension and recognition of study information, is an important element of informed consent. Research suggests that current standards and practices for consent procedures do not adequately promote ideal participant understanding, specifically in vulnerable populations that may have reduced decision-making capacities (Morán-Sánchez et al., 2016; Neilson et al., 2015; Westra & de Beaufort, 2015). Therefore, individuals may not be able to completely comprehend or recall critical aspects of the study procedures, risks, benefits, and human subject protections (Festinger et al., 2007; Madeira & Andraka-Christou, 2016).

Therapeutic Misconception

When making the decision to participate in research, some individuals may falsely believe that any treatment provided in the context of research has therapeutic value. Of course, research may include placebo conditions with no therapeutic value and novel interventions may lack efficacy. These false beliefs, referred to as therapeutic misconception (Appelbaum et al., 1982), commonly present in three ways:

(1) a false belief that treatment provided by the research will be tailored to a participants' needs
(2) an inability to discern that the principal aim of the research is to progress scientific knowledge and not to specifically benefit participants
(3) an unrealistic expectation of therapeutic benefits (Christopher et al., 2016).

It is important to consider the potential for therapeutic misconceptions in each study as they may prevent individuals from accurately evaluating the study's risks and benefits (Appelbaum et al., 1982; Christopher et al., 2016; Lidz, 2006). Assuring that participants completely comprehend the differences between treatments received in research and alternatives to standard treatment, by means of education during the consent process, can significantly reduce the potential for therapeutic misconception (see Christopher et al., 2017).

Understanding and Recalling Informed Consent and Study Information

Research participants must have a strong comprehension of the potential risks and benefits of their participation and understand the difference between research and treatment. However, studies have indicated that research participants often do not recognize that they are taking part in research, have limited comprehension of study information, do not understand the risks and benefits of participation, do not comprehend the concept of randomization or placebo/control conditions, and are unaware that, at any time, they can choose to withdraw from the study (e.g., Appelbaum et al., 1982; Edlund et al., 2015; Festinger et al., 2007, 2009). Additionally, only a few days after having provided consent, participants often fail to retain much of the information presented at the time of consent (e.g., Festinger et al., 2007, 2009; Miller et al., 1994; Rosique et al., 2006). These findings challenge the consent process as informed and indicate that the consent process should be modified to ensure comprehension and retention of the information presented during the consent process.

Strategies for Improving Knowingness

Several effective strategies exist to improve comprehension and retention of information presented during the informed consent process and include either changing the structure of the consent process or altering the process itself.

Modifying the Structure of Consent Documents

Several strategies focused on consent-related documents are effective in bolstering comprehension and recall of information. Generally, consent documents are written at an adult reading level or, in the case of children, at their specific grade level. Improving the readability of consent forms is imperative given that approximately

21% of adults in the United States have marginal literacy skills (US Department of Education, 2019). However, it is often difficult for individuals to comprehend and retain the most important consent information even when consent documents utilize appropriate reading levels (e.g., Muir & Lee, 2009).

When considering potential modifications to the structure of a consent form, shortening the form and using more succinct language are often more effective at relaying pertinent information and heightening understanding than lengthier consent forms (Beardsley et al., 2007; Enama et al., 2012; Matsui et al., 2012; Stunkel et al., 2010). A meta-analysis showed that using simpler language and revising layouts, text styling, and diagrams is effective at significantly improving the participants' understanding of information presented in the consent process (Nishimura et al., 2013). In a more recent study, Kim and Kim (2015) found that a streamlined consent form with larger font, wider spacing, shorter sentences, and the inclusion of pictures, diagrams, bulleting, and clearer text styling was effective at improving several aspects of the participants' objective and subjective understanding of the consent and study information, such as the length of study participation, study procedures, randomization, alternatives to standard treatment, risks and reimbursement for harm, and participants' freedom to revisit the consent document at any time.

Improving the Consent Process

Corrected feedback has received the most empirical support as a consent strategy. It involves evaluating a potential participant's understanding of consent information, following a review of the consent form, and then addressing incorrect responses with the participant to ensure they know the correct answer. Corrected feedback improves both initial comprehension, at the time of consent, and longer-term recall of consent information (Carpenter et al., 2000; Coletti et al., 2003; Festinger et al., 2010; Stiles et al., 2001; Taub & Baker, 1983; Taub et al., 1981; Wirshing et al., 1998). Additionally, comprehension can be further improved by linking additional structural interventions, especially related to randomization, with corrected feedback procedures (Kass et al., 2015).

Consent Quizzes

Consent quizzes provide an objective way to determine the extent to which an individual understands consent information. Furthermore, they can help to identify areas that require clarification (Allen et al., 2017). Although researchers generally agree on the importance of using assessments to gauge understanding of consent materials (e.g., Appelbaum & Grisso, 2001; Edlund et al., 2015; Festinger et al., 2009, 2014), consent quizzes often assess recognition rather than recall. This is problematic, as it does not reflect how this information would be used by participants in the real world where they would have to actually *recall* study-related information rather than rely on recognition memory. Table 10.1 provides examples of commonly used multiple choice and true/false questions along with suggestions about how they can be transformed into open-ended questions to better evaluate understanding and recall.

Table 10.1 *Informed consent quiz best practices*

Original multiple-choice item	Improved open-ended item
How many times will you complete interviews in this study? (a) 3 times; (b) 4 times; (c) 5 times; (d) 6 times.	How many interviews will you be asked to complete in this study?
True or false: I can leave the study at any time.	What happens if you don't want to be in the study anymore?
How much will you be paid to be in the study? (a) \$20; (b) \$30; (c) \$40; (d) \$50.	What amount of payment will you receive for being in the study?
True or false: The two groups in the study are the control and the intervention.	How many different groups are there in this study? How do the groups differ?

Strengthening Motivation

Although the remedial process of modifying the informed consent structure and procedures can be effective at improving comprehension and recall of the information presented in the informed consent process, several other factors may impact participants' comprehension and recall. Several variables that impact cognition (e.g., level of education, neuropsychological measures of memory and attention, and IQ) are related to an individual's ability to accurately recall consent information (Dunn & Jeste, 2001; Flory & Emanuel, 2004; Taub & Baker, 1983; Taub et al., 1981); however, these cognitive variables account for less than half of the variance in recall (Festinger et al., 2007). This suggests that remedial strategies designed to simplify the cognitive task do not fully address the issue.

Some individuals might not understand or recall consent information because they do not adequately attend to and/or process information presented during the consent process. Festinger et al. (2009) manipulated the motivation to attend to consent information by offering incentives of \$5 for each consent quiz question that they correctly answered, one week following the initial date of consent. Those who were offered incentives displayed increases in recall of consent information relative to those who did not, emphasizing the role of motivation in the consent process. Furthermore, recall of consent information increased when incentives were combined with remedial corrected feedback procedures (Festinger et al., 2014). Although this may not be a practical strategy in all study contexts, it provides evidence for the role of motivation in consent recall.

Multimedia Consent Approaches

Researchers continue to develop ways to incorporate multimedia methods and nontraditional modalities (e.g., video, computer, mobile phone, and web-based applications) into the informed consent process. A meta-analysis of studies using multimedia consent approaches conducted by Nishimura et al. (2013) found 31% of studies reviewed reported significant gains in understanding, and several studies demonstrated significant increases in the retention and recall of study-related information.

Several studies demonstrated that videos with audio narration can improve participants' comprehension (Kraft et al., 2017; Rothwell et al., 2014; Spencer et al., 2015; Winter et al., 2016) and retention and recall of consent and study information (Siu et al., 2016; Tipotsch-Maca et al., 2016). Sonne and colleagues (2013) found that most research participants preferred this modality relative to standard paper-and-pencil consent forms; participants reported that the video format assisted their decision-making and understanding of study procedures. Nevertheless, the findings in this area appear mixed and may very well be moderated by the type of study, patient population, and other factors (e.g., Frost et al., 2021; Rothwell et al., 2014). Additionally, the use of animated videos without an audio component improves comprehension of consent material; however, participants using this approach reported lower levels of satisfaction (Bowers et al., 2017; Ham et al., 2016). Importantly, these various multimedia consent procedures are particularly useful for certain types of individuals, such as those with low levels of literacy (Afolabi et al., 2014, 2015), diminished capacity (Morán-Sánchez et al., 2016), and psychiatric comorbidities (Jeste et al., 2009; Sonne et al., 2013).

Obtaining Informed Consent Remotely

Healthcare providers have been increasingly incorporating telehealth strategies to provide remote care to patients. The use of telemedicine in medical practice has improved the accessibility of care for patients, especially in rural and underserved populations, while reducing costs and maintaining patient satisfaction (Welch et al., 2016). Similar approaches may have value for consenting and conducting research; however, few empirical studies have tested their efficacy. Bobb et al. (2016) tested the efficacy of a remote consent procedure compared to standard in-person consent and found no differences in the two procedures in comprehension or recruitment rates, indicating that the use of remote consent procedures does not diminish potential participants' understanding of consent information or willingness to participate compared to standard consent. Additionally, multimedia consent approaches are beneficial in research involving children and adolescents, where assent may be obtained in addition to parent or guardian consent (Martin-Kerry et al., 2017; Sheridan et al., 2019). Although these research findings are promising, additional research needs to assess the utility and cost-effectiveness associated with remote procedures.

As access to technology continues to expand, with approximately 89% of US households owning computers or smartphones (Ryan, 2018), multimedia approaches to informed consent will likely become standard practice. Despite general findings supporting the use of multimedia consent, it is difficult to draw definitive conclusions regarding effectiveness given the heterogeneity in consent procedures, study complexity, and targeted populations.

Voluntariness

The decision to participate in research must be made autonomously without any coercion or undue influence. These two constructs are often used interchangeably, but their definitions are somewhat distinct. Coercion occurs when a person of authority implicitly or explicitly threatens some form of harm to the individual if they do not behave in a certain way (e.g., participate in a study). Undue influence occurs when something of exceptionally high value is offered in return for a person engaging in a specific behavior. Simple remuneration or positive reinforcement does not by itself constitute undue influence; the offer must be of such great value to the individual that it throws off the individual's ability to refuse. For example, remuneration of $1,000 may be deemed by an IRB to be undue influence in a study of persons of lower socio-economic status while it may not represent undue influence in a more affluent sample. Importantly, IRBs must always consider the risk/benefit ratio of studies when making these determinations.

Participant characteristics and circumstances may present challenges to autonomous decision-making. It is conceivable that participants in lower-income situations would accept a higher level of risk for financial gain. Similarly, individuals who are incarcerated may participate in research to avoid fear of punishment (i.e., coercion) or to reduce time behind bars (i.e., undue influence). Importantly, risks to voluntariness can be real or perceived – research participants regularly report feeling pressured when they have, in fact, not been pressured at all (Appelbaum et al., 2009).

Strategies to Promote Voluntariness

Assessment

There are several validated assessments that can be used to determine whether an individual's decision to enroll in a study is autonomous and free from undue influence and coercion. The MacArthur Perceived Coercion scale (PCS; Gardner et al., 1993) is a brief, five-item true/false measure intended to measure a patient's feelings of coercion related to inpatient psychiatric hospitalization. Appelbaum et al. (2009) modified the instrument for use as a measure of research-related coercion (e.g., Festinger et al., 2008; Moser et al., 2004). Similarly, the Iowa Coercion Questionnaire (ICQ) builds upon the PCS by including items that assess self-presentation concerns (e.g., "I entered the study to appear cooperative"; Moser et al, 2004). Although these measures show utility in assessing individuals' feelings of pressure or coercion, they do not identify the source or magnitude of coercion or undue influence. The Coercion Assessment Scale (CAS; Dugosh et al., 2010, 2014) was developed to address these limitations. In addition to measuring the presence of coercion and undue influence, the instrument identifies their source and magnitude. This level of specificity allows researchers to take actionable steps to increase voluntariness (see Box 10.2).

Box 10.2 Coercion Assessment Scale (Dugosh et al., 2010, 2014)

Participants are asked to rate the veracity of each item on a four-point scale:

(1) I felt like I was talked into entering the study.
(2) It was entirely my choice to enter the study.
(3) I thought that it would look bad to my *healthcare provider* if I did not enter the study.
(4) I felt like my *healthcare provider* would like it if I entered the study.
(5) I entered the study even though I did not want to.
(6) I entered the study mainly for financial reasons.
(7) I felt that I could not say "no" to entering the study.
(8) I felt that entering the study would help me get better medical care.

Any item responses suggesting that a participant may be feeling coerced or unduly influenced should be discussed and addressed with the participant.

Research Intermediaries

Research intermediaries (i.e., research advocates, ombudsmen, and neutral educators) can help promote the autonomy of research participants by assisting them in making informed decisions (Benson, et al., 1988; Reiser & Knudson, 1993; Stroup & Appelbaum, 2003). They are independent and have no ties to the research being conducted or related entities and can be other patients, caregivers, or other staff members or trained professionals. Research intermediaries can explain study protocols, including the risks and benefits of research participation, and act as an advocate for the participant. Numerous studies have shown that research intermediaries enhance comprehension and recall of consent information (e.g., Coletti et al., 2003; Fitzgerald, et al., 2002; Kucia & Horowitz, 2000) and mitigate undue influence and perceived coercion (Festinger, et al., 2011). Notably, the use of research intermediaries has been endorsed by several federal advisory panels and agencies to enhance the informed consent process. (e.g., National Bioethics Advisory Commission, 1998; National Institutes of Health, 2009; World Medical Association, 2013).

Conclusion

The informed consent process continues to evolve as the foundation for ethical research. It is an ongoing process rather than a one-time event and is designed to ensure that research participants (1) have the capacity to make decisions about engaging in research, (2) are fully informed, understand, and recall the research process, the potential risks and harms of participation, and their protections, and (3) do so voluntarily. A substantial body of research has identified numerous evidence-based strategies to improve the effectiveness of the consent procedure. Despite these many advances, empirical research on improving human subjects protections must continue. Moreover, it is critically important to translate these empirical findings into practice and policy.

References

Afolabi, M. O., Bojang, K., D'Alessandro, U., et al. (2014). Multimedia informed consent tool for a low literacy African research population: Development and pilot testing. *Journal of Clinical Research and Bioethics*, *5*(3), 178.

Afolabi, M. O., McGrath, N., D'Alessandro, U., et al. (2015). A multimedia consent tool for research participants in the Gambia: A randomized controlled trial. *Bulletin of the World Health Organization*, *93*(5), 320–328A.

Allen, A. A., Chen, D. T., Bonnie, R. J., et al. (2017). Assessing informed consent in an opioid relapse prevention study with adults under current or recent criminal justice supervision. *Journal of Substance Abuse Treatment*, *81*, 66–72.

Appelbaum, P. S. & Grisso, T. (2001). *Macarthur Competence Assessment Tool for Clinical Research (MacCAT-CR)*. Professional Resource Press/Professional Resource Exchange.

Appelbaum, P. S. & Grisso, T. (2007) Assessment of patients' competence to consent to treatment. *The New England Journal of Medicine*, *357*(18), 1834–1840.

Appelbaum, P. S., Roth, L. H., & Lidz, C. (1982). The therapeutic misconception: Informed consent in psychiatric research. *International Journal of Law and Psychiatry*, *5*, 319–329.

Appelbaum, P. S., Lidz, C. W., & Klitzman, R. (2009). Voluntariness of consent to research: A conceptual model. *Hastings Center Report*, *39*(1), 30–39.

Beardsley, E., Jefford, M., & Mileshkin, I. (2007). Longer consent forms for clinical trials compromise patient understanding: So why are they lengthening? *Journal of Clinical Oncology*, *23*(9), e13–e14.

Benson, P. R., Roth, L. H., Appelbaum, P. S., Lidz, C. W., & Winslade, W. J. (1988). Information disclosure, subject understanding, and informed consent in psychiatric research. *Law and Human Behavior*, *12*(4), 455–475.

Bobb, M. R., Van Heukelom, P. G., Faine, B. A., et al. (2016). Telemedicine provides noninferior research informed consent for remote study enrollment: A randomized controlled trial. *Academic Emergency Medicine*, *23*(7), 759–765.

Bowers, N., Eisenberg, E., Montbriand, J., Jaskolka, J., & Roche-Nagle, G. (2017). Using a multimedia presentation to improve patient understanding and satisfaction with informed consent for minimally invasive vascular procedures. *The Surgeon*, *15*(1), 7–11.

California Legislative Information (2003). Human experimentation, California Health and Safety Code – HSC § 24178. Available at: https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?lawCode=HSC&division=20.&title=&part=&chapter=1.3.&article.

Carlson, L. (2013). Research ethics and intellectual disability: Broadening the debate. *Yale Journal of Biology and Medicine*, *86*, 303–314.

Carpenter Jr., W. T., Gold, J. M., Lahti, A. C., et al. (2000). Decisional capacity for informed consent in schizophrenia research. *Archives of General Psychiatry*, *57*(6), 533–538.

Christopher, P. P., Stein, M. D., Springer, S. A., et al. (2016). An exploratory study of therapeutic misconception among incarcerated clinical trial participants. *AJOB Empirical Bioethics*, *7*(1), 24–30.

Christopher, P. P., Appelbaum, P. S., Truong, D., et al. (2017). Reducing therapeutic misconception: A randomized intervention trial in hypothetical clinical trials. *PLoS ONE*, *12*(9), e018224.

Coletti, A. S., Heagerty, P., Sheon, A. R., et al. (2003). Randomized, controlled evaluation of a prototype informed consent process for HIV vaccine efficacy trials. *Journal of Acquired Immune Deficiency Syndromes*, *32*(2), 161–169.

Davies, A. N., Waghorn, M., Webber, K., et al. (2018). A cluster randomised feasibility trial of clinically assisted hydration in cancer patients in the last days of life. *Palliative Medicine*, *32*(4), 733–743.

De Sutter, E., Zace, D., Boccia, S., et al. (2020) Implementation of electronic informed consent in biomedical research and stakeholders" perspectives: Systematic review. *Journal of Medical Internet Research*, *22*(10), e19129.

Dugosh, K. L., Festinger, D. S., Croft, J. R., & Marlowe, D. B. (2010). Measuring coercion to participate in research within a doubly vulnerable population: Initial development of the coercion assessment scale. *Journal of Empirical Research on Human Ethics*, *5* (1), 93–102.

Dugosh, K. L., Festinger, D. S., Marlowe, D. B., & Clements, N. T. (2014). Developing an index to measure the voluntariness of consent to research. *Journal of Empirical Research on Human Ethics*, *9*(4), 60–70.

Dunn, L. B. & Jeste, D. V. (2001). Enhancing informed consent for research and treatment. *Neuropsychopharmacology*, *24*(6), 595–607.

Dunn, L. B., Nowrangi, M. A., Palmer, B. W., Jeste, D. V., & Saks, E. R. (2006). Assessing decisional capacity for clinical research or treatment: A review of instruments. *American Journal of Psychiatry*, *163*(8), 1323–1334.

Edlund, J. E., Edlund, A. E., & Carey, M. G. (2015). Patient understanding of potential risk and benefit with informed consent in a left ventricular assist device population: A pilot study. *Journal of Cardiovascular Nursing*, *30*(5), 435–439.

Electronic Code of Federal Regulations (2018). Protection of Human Subjects, 45 C.F.R. § 46. Available at: www.ecfr.gov/on/2018-07-19/title-45/subtitle-A/subchapter-A/part-46.

Enama, M. E., Hu, Z., Gordon, I., et al. (2012). Randomization to standard and concise informed consent forms: Development of evidence-based consent practices. *Contemporary Clinical Trials*, *33*, 895–902.

Evans, C. J., Yorganci, E., Lewis, P., et al. (2020). Processes of consent in research for adults with impaired mental capacity nearing the end of life: Systematic review and transparent expert consultation (MORECare_Capacity statement). *BMC Medicine*, *18*(1), 221.

Faden, R. (1996). The Advisory Committee on Human Radiation Experiments. *Hastings Center Report*, *26*(5), 5–10.

Festinger, D. S., Ratanadilok, K., Marlowe, D. B., et al. (2007). Neuropsychological functioning and recall of research consent information among drug court clients. *Ethics & Behavior*, *17*(2), 163–186.

Festinger, D., Marlowe, D., Dugosh, K., Croft, J., & Arabia, P. (2008). Higher magnitude cash payments improve research follow-up rates without increasing drug use or perceived coercion. *Drug and Alcohol Dependence*, *96*(1–2), 128–135.

Festinger, D. S., Marlowe, D. B., Croft, J. R., et al. (2009). Monetary incentives improve recall of research consent information: It pays to remember. *Experimental and Clinical Psychopharmacology*, *17*(2), 99–104.

Festinger, D. S., Dugosh, K. L., Croft, J. R., Arabia, P. L., & Marlowe, D. B. (2010). Corrected feedback: A procedure to enhance recall of informed consent to research among substance abusing offenders. *Ethics & Behavior*, *20*(5), 387–399.

Festinger, D. S., Dugosh, K. L., Croft, J. R., Arabia, P. L., & Marlowe, D. B. (2011). Do research intermediaries reduce perceived coercion to enter research trials among criminally involved substance abusers? *Ethics & Behavior*, *21*(3), 252–259.

Festinger, D. S., Dugosh, K. L., Marlowe, D. B., & Clements, N. (2014). Achieving new levels of recall in consent to research by combining remedial and motivational techniques. *Journal of Medical Ethics*, *40*(4), 264–268.

Fitzgerald, D. W., Marotte, C., Verdier, R. I., Johnson, W. D., & Pape, J. W. (2002). Comprehension during informed consent in a less-developed country. *Lancet*, *360*, 1301–1302.

Flory, J. & Emanuel, E. (2004). Interventions to improve research participants' understanding in informed consent for research: A systematic review. *Journal of the American Medical Association*, *292*, 1593–1601.

Forster, D. G. & Borasky, D. A., Jr. (2018). Adults lacking capacity to give consent: When is it acceptable to include them in research? *Therapeutic Innovation & Regulatory Science*, *52*(3), 275–279.

Frost, C. J., Johnson, E. P., Witte, B., et al. (2021). Electronic informed consent information for residual newborn specimen research: Findings from focus groups with diverse populations. *Journal of Community Genetics*, *12*(1), 199–203.

Gardner, W., Hoge, S. K., Bennett, N., et al. (1993). Two scales for measuring patients' perceptions for coercion during mental hospital admission. *Behavioral Sciences and the Law*, *11*(3), 307–321.

Gilbert, T., Bosquet, A., Thomas-Anterion, C., Bonnefoy, M., & Le Saux, O. (2017). Assessing capacity to consent for research in cognitively impaired older patients. *Clinical Interventions in Aging*, *12*, 1553–1563.

Gupta, U. C. (2013). Informed consent in clinical research: Revisiting few concepts and areas. *Perspectives in Clinical Research*, *4*(1), 26–32.

Ham, D. Y., Choi, W. S., Song, S. H., et al. (2016). Prospective randomized controlled study on the efficacy of multimedia informed consent for patients scheduled to undergo green-light high-performance system photoselective vaporization of the prostate. *World Journal of Men's Health*, *34*(1), 47–55.

Ho, P., Downs, J., Bulsara, C., Patman, S., & Hill, A. (2017) Addressing challenges in gaining informed consent for a research study investigating falls in people with intellectual disability. *British Journal of Learning Disabilities*, *46*(2), 92–100.

International Military Tribunal (1950). *Trials of War Criminals before the Nuremberg Military Tribunals under Control Council Law No. 10*. Government Printing Office.

Jeste, D. V., Palmer, B. W., Appelbaum, P. S., et al. (2007). A new brief instrument for assessing decisional capacity for clinical research. *Archives of General Psychiatry*, *64*(8), 966–974.

Jeste, D. V., Palmer, B. W., Golshan, S., et al. (2009). Multimedia consent for research in people with schizophrenia and normal subjects: A randomized controlled trial. *Schizophrenia Bulletin*, *35*(4), 719–729.

Jones, J. H. (1993). *Bad Blood: The Tuskegee Syphilis Experiment*. The Free Press.

Kass, N., Taylor, H., Ali, J., Hallez, K. & Chaisson, L. (2015). A pilot study of simple interventions to improve informed consent in clinical research: Feasibility, approach, and results. *Clinical Trials*, *12*(1), 54–66.

Kim, E. J. & Kim, S. H. (2015). Simplification improves understanding of informed consent information in clinical trials regardless of health literacy level. *Clinical Trials*, *12* (3), 232–236.

Kraft, S. A., Constantine, M., Magnus, D., et al. (2017). A randomized study of multimedia informational aids for research on medical practices: Implications for informed consent. *Clinical Trials*, *14*(1), 94–102.

Kucia, A. M. & Horowitz, J. D. (2000). Is informed consent to clinical trials and "upside selective" process in acute coronary syndromes. *American Heart Journal*, *140*, 94–97.

Layman, E. (2009). Human experimentation: Historical perspective of breaches of ethics in U.S. healthcare. *The Health Care Manager*, *28*(4), 354–374.

Lepola, P., Needham, A., Mendum, J., et al. (2016). Informed consent for paediatric clinical trials in Europe. *Archives of Disease in Childhood*, *101*(11), 1017–1025.

Lidz, C. W. (2006). The therapeutic misconception and our models of competency and informed consent. *Behavioral Sciences & the Law*, *24*(4), 535–546.

Madeira, J. L. & Andraka-Christou, B. (2016). Paper trials, trailing behind: Improving informed consent to IVF through multimedia approaches. *Journal of Law and the Biosciences*, *3*(1), 2–38.

Maloy, J. W. & Bass, P. F. (2020). Understanding broad consent. *Ochsner Journal*, *20*(1), 81–86.

Martel, M. L., Klein, L. R., Miner, J. R., et al. (2018). A brief assessment of capacity to consent instrument in acutely intoxicated emergency department patients. *American Journal of Emergency Medicine*, *36*, 18–23.

Martin-Kerry, J., Bower, P., Young, B., et al. (2017). Developing and evaluating multimedia information resources to improve engagement of children, adolescents, and their parents with trials (TRECA study): Study protocol for a series of linked randomised controlled trials. *Trials*, 18, 265.

Matsui, K., Lie, R. K., Turin, T. C., & Kita, Y. (2012). A randomized controlled trial of short and standard-length consent for a genetic cohort study: Is longer better? *Journal of Epidemiology*, *22*, 308–316.

Milgram, S. (1974). *Obedience to Authority*. Harper Collins.

Miller, C. M., Searight, H. R., Grable, D., et al. (1994). Comprehension and recall of the informational content of the informed consent document: An evaluation of 168 patients in a controlled clinical trial. *Journal of Clinical Research and Drug Development*, *8*(4), 237–248.

Morán-Sánchez, I., Luna, A., & Pérez-Cárceles, M. D. (2016). Enhancing the informed consent process in psychiatric outpatients with a brief computer-based method. *Psychiatry Research*, *245*, 354–360.

Moser, D. J., Arndt, S., Kanz, J. E., et al. (2004). Coercion and informed consent in research involving prisoners. *Comprehensive Psychiatry*, *45*(1), 1–9.

Muir, K. W. & Lee, P. P. (2009). Literacy and informed consent: A case for literacy screening in glaucoma research. *Archives of Ophthalmology*, 127(5), 698–699

National Bioethics Advisory Commission. (1998). Research involving persons with mental disorders that may affect decision making capacity. Available at: https://bioethicsarchive.georgetown.edu/nbac/capacity/TOC.htm.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (1979). *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. US Department of Health and Human Services.

National Institutes of Health (2009). Research involving individuals with questionable capacity to consent: Points to consider. Available at: https://grants.nih.gov/grants/policy/questionablecapacity.htm.

Neilson, G., Chaimowitz, G., & Zuckerberg, J. (2015). Informed consent to treatment in psychiatry. *Canadian Journal of Psychiatry*, *60*, 1–12.

Nishimura, A., Carey, J., Erwin, P. J., et al. (2013). Improving understanding in the research informed consent process: A systematic review of 54 interventions tested in randomized control trials. *BMC Medical Ethics*, *14*, 28.

Palmer, B. W., Dunn, L. B., Appelbaum, P. S., et al. (2005). Assessment of capacity to consent to research among older persons with schizophrenia, Alzheimer's disease, or diabetes mellitus. *Archives of General Psychiatry*, *62*(7), 726–733.

Reiser, S. J. & Knudson, P. (1993). Protecting research subjects after consent: The case for the research intermediary. *IRB*, *15*(2), 10–11.

Rosique, I., Pérez-Cárceles, M. D., Romero-Martín, M., Osuna, E., & Luna, A. (2006). The use and usefulness of information for patients undergoing anaesthesia. *Medicine and Law*, *25*(4), 715–727.

Rothwell, E., Wong, B., Rose, N. C., et al. (2014). A randomized controlled trial of an electronic informed consent process. *Journal of Empirical Research on Human Research Ethics*, *9*(5), 1–7.

Ryan, C. (2018). Computer and Internet use in the United States: 2016. American Community Survey Reports. Available at: www.census.gov/library/publications/2018/acs/acs-39.html.

Seaman, J.B., Terhorst, L., Gentry, A., et al. (2015). Psychometric properties of a decisional capacity screening tool for individuals contemplating participation in Alzheimer's disease research. *Journal of Alzheimer's Disease*, *46*(1), 1–9.

Sheridan, R., Martin-Kerry, J., Watt, I., et al. (2019). User testing digital, multimedia information to inform children, adolescents and their parents about healthcare trials. *Journal of Child Health Care*, *23*(3), 468–482.

Shuster, E. (1997). Fifty years later: The significance of the Nuremberg Code. *The New England Journal of Medicine*, *337*(20), 1436–1440.

Siu, J. M., Rotenberg, B. W., Franklin, J. H., & Sowerby, L. J. (2016). Multimedia in the informed consent process for endoscopic sinus surgery: A randomized control trial. *Laryngoscope*, *126*(6), 1273–1278.

Sonne, S. C., Andrews, J. O., Gentilin, S. M., et al. (2013). Development and pilot testing of a video-assisted informed consent process. *Contemporary Clinical Trials*, *36*(1), 25–31.

Spencer, S. P., Stoner, M. J., Kelleher, K., & Cohen, D. M. (2015). Using a multimedia presentation to enhance informed consent in a pediatric emergency department. *Pediatric Emergency Care*, *31*(8), 572–576.

Stiles, P. G., Poythress, N. G., Hall, A., Falkenbach, D., & Williams, R. (2001). Improving understanding of research content disclosures among persons with mental illness. *Psychiatric Services*, *52*, 780–785.

Stroup, S. & Appelbaum, P. (2003). The subject advocate: Protecting the interests of participants with fluctuating decision making capacity. *IRB*, *25*(3), 9–11.

Stunkel, L., Benson, M., McLellan, L., et al. (2010). Comprehension and informed consent: Assessing the effect of a short consent form. *IRB: Ethics & Human Research*, *32*(4), 1–9.

Taub, H. A. & Baker, M. T. (1983). The effect of repeated testing upon comprehension of informed consent materials by elderly volunteers. *Experimental Aging Research*, *9*, 135–138.

Taub, H. A., Kline, G. E., & Baker, M. T. (1981). The elderly and informed consent: Effects of vocabulary level and corrected feedback. *Experimental Aging Research*, *7*, 137–146.

Tipotsch-Maca, S. M., Varsits, R. M., Ginzel, C., & Vescei-Marlovits, P. V. (2016). Effect of a multimedia-assisted informed consent procedure on the information gain, satisfaction, and anxiety of cataract surgery patients. *Journal of Cataract and Refractive Surgery*, *42*(1), 110–116.

US Department of Education (2019). Adult literacy in the United States. Available at: https://nces.ed.gov/pubs2019/2019179/index.asp.

US Food and Drug Administration (2016). Use of electronic informed consent: Questions and answers. Available at: www.fda.gov/regulatory-information/search-fda-guidance-documents/use-electronic-informed-consent-clinical-investigations-questions-and-answers.

Welch, B. M., Marshall, E., Qanungo, S., et al. (2016). Teleconsent: A novel approach to obtain informed consent for research. *Contemporary Clinical Trials Communications*, *15*(3), 74–79.

Westra, A. E. & de Beaufort, I. (2015). Improving the Helsinki Declaration's guidance on research in incompetent subjects. *Journal of Medical Ethics*, *41*, 278–280.

Winter, M., Kam, J., Nalavenkata, S., et al. (2016). The use of portable video media vs standard verbal communication in the urological consent process: A multicenter, randomised controlled, crossover trial. *BJU International*, *118*(5), 823–828.

Wirshing, D. A., Wirshing, W. C., Marder, S. R., Liberman, R. P., & Mintz, J. (1998). Informed consent: Assessment of comprehension. *American Journal of Psychiatry*, *155*, 1508–1511.

World Medical Association (2013). World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. *Journal of the American Medical Association*, *310*, 2191–2194.

11 Experimenter Effects

Jocelyn Parong, Mariya Vodyanyk, C. Shawn Green, Susanne M. Jaeggi, and Aaron R. Seitz

Abstract

As social and behavioral scientists, it is of fundamental importance to understand the factors that drive the behaviors that we measure. Careful design is thus required to minimize the influence of extraneous factors. Yet, we often overlook one major class of such extraneous factors – those related to us, the experimenters. Experimenter effects can potentially arise at every step in the research process – from the selection of hypotheses, to interacting with research participants in ways that might alter their behavior, to biases in data interpretation. While such experimenter-driven effects often occur without notice, and without ill intent, they nonetheless threaten the replicability and generalizability of research. In this chapter, we discuss when and how such effects arise, preventative measures that can be taken to reduce their influence, and methods for accounting for such effects, when appropriate.

Keywords: Experimenter Effect, Expectancy Effect, Observer Effect, *p*-Hacking, HARK-ing

Introduction

In the late 1800s, a horse trainer named Wilhelm von Osten wowed crowds by demonstrating that his horse had almost human-like intelligence. Von Osten would ask his horse, aptly named Clever Hans, a wide variety of questions, such as those involving basic arithmetic, identifying colors, or even reading and spelling words, and Hans would correctly respond with a series of hoof taps. For example, when asked, "If the eighth day of the month comes on a Tuesday, what is the date of the following Friday?," Hans would tap his hoof 11 times and then stop. Was this horse capable of comprehending these complicated questions or was there another logical explanation?

In 1907, psychologist Oskar Pfungst evaluated von Osten and Hans. Interestingly, he found that Hans only responded correctly when the questioners knew the answer themselves, whether the questioner was von Osten or someone else. After examining the questioners' behaviors, Pfungst discovered that Hans reacted to (potentially) unintentional body language, such as posture or facial expression, that questioners exhibited after Hans performed the correct number of hoof taps (Pfungst, 1911). Though unnamed at the time, the Clever Hans phenomenon became one of the earliest examples of an *experimenter effect* – the influence exerted by experimenters on experimental outcomes (Rosenzweig, 1933).

Over the subsequent century, experimenter effects have been studied extensively in research (see Rosenthal, 1997 for a review). Research has clearly demonstrated that experimenters unintentionally exert influence on essentially every aspect of the research process – the formulation of hypotheses, aspects of study designs, interactions with participants during data collection, decisions regarding analyses, and the conclusions that are drawn. Researchers enter projects with certain beliefs, attitudes, expectations, and knowledge. Critically, these prior beliefs influence the research and can threaten replicability and generalizability of the research (Edlund et al., 2021). Indeed, if study outcomes are driven by experimenter effects, similar outcomes may not be found by a group that enters a replication with a different set of beliefs and biases.

This chapter reviews various types of experimenter effects that can occur in three broad steps of the research process: study design, data collection, and data analysis and interpretation. Implications for experimenter effects in research, including how to minimize them (or take advantage of them, when appropriate) and/or how to account for them in situations where minimizing such effects is either not possible or not desirable are also discussed.

Experimenter Effects During Study Design

The first step in designing a study is typically gathering information about previous research in the field. Understanding past research methods and findings allows us to identify gaps in knowledge and paths to discovery. While one might imagine that reviewing previous research offers little opportunity for bias, in practice, the opposite is true.

For example, *confirmation bias* refers to the well-known tendency to seek out evidence that supports our beliefs while also dismissing contradictory evidence (Nickerson, 1998). In the classic example, you are shown four cards, each with a number on one side and a color on the other side. The visible side of the cards show the number 3, the number 8, the color red, and the color blue, respectively. You are then given the hypothesis, "If an even number appears on one side of a card, then the opposite side is red." Which cards do you need to flip over to test this hypothesis? If you chose to flip over the card with the color red showing, you, like the majority of responders on this task, have displayed confirmation bias (Wason, 1968). You chose a card that could support the hypothesis but could not disconfirm it. That is, if an even number appears on the opposite side of the red card, it supports the hypothesis, but if an odd number appears, it does not disconfirm it. As such, you were biased to find evidence in support of your hypothesis rather than searching for evidence that could disprove it. Fewer than 10% of participants correctly identified the two cards that could disconfirm the hypothesis – the number 8 (if a color other than red appeared on the opposite side) and the blue card (if an even number appeared on the opposite side).

Because people tend to prefer information that confirms their pre-existing attitudes, they may selectively search for information that conforms to those attitudes (*selective exposure;* Hart et al., 2009). A meta-analysis found that people are twice as likely to select information that agrees, rather than disagrees, with their pre-existing attitudes, beliefs, and behaviors (Hart et al., 2009). This may be particularly troublesome when evidence is mixed or inconclusive, as experimenters may not incorporate contradictory information into their broader considerations of the existing state of scientific knowledge.

As one example, researchers' attitudes toward video games may determine how they choose to conduct research on the possible impact of playing video games. A researcher starting from the perspective that video games are harmful is more likely to utilize a psychopathological lens and search for research demonstrating problematic outcomes from video games. Meanwhile, a researcher starting with a positive perspective about playing video games is more likely to view video game research as a tool for improving psychological health and well-being and thus search for papers consistent with this view (Klecka et al., 2021; Meier, et al., 2020). For a further discussion, see Chapter 4 in this volume.

Confirmation bias can also manifest in selection of methods that are more likely to support, and less likely to disconfirm, pre-existing beliefs or expectations. Indeed, design decisions regarding any part of the study, including the design of stimuli, types of control or comparison groups, and/or the specific procedures that are used, can impact participant behavior. Importantly, it is often possible for researchers to intuit how certain differences in procedures, for instance, might impact participant behavior. In this vein, Forster (2000) examined whether researchers could intuit which stimuli in a lexical decision-making task could better support a hypothesis. In this task, participants are presented with an English word and a non-word of similar length and must decide which is an English word. When researchers were given pairs of the English words from the task and were asked to identify which word in the pair would prompt a faster reaction time, they were able to reliably do so.

Critically, this ability to anticipate how participants will respond, if unchecked, can result in the use of methods that are more likely to match pre-existing beliefs. For example, Strickland and Suben (2012) asked research assistants to develop sentences to test a hypothesis about whether using certain "non-feeling" (e.g., intend, want) or "feeling" (e.g., experience, suffer) words in a sentence would be rated by participants as more natural sounding. Some research assistants were told that the hypothesis was that sentences with "feeling" words would sound more natural, while others were told the opposite. Participants were then asked to rate how natural these sentences sounded. When research assistants were told that the hypothesis was that "feeling" words would sound more natural, participants rated those sentences with "feeling" words as more natural than the "non-feeling" sentences. The opposite was true for research assistants who were told that "non-feeling" words would sound more natural. That is, the research assistants, either intentionally or unintentionally, wrote sentences in line with their given hypothesis about the study. Thus, it is important to consider how an experimenter's prior beliefs or expectations influence the design of stimulus material.

Experimenter Effects During Data Collection

During data collection, experimenter effects can be broadly grouped into (a) expectancy effects, (b) participant reactivity effects, and (c) observer effects. *Expectancy effects* describe situations where the experimenter’s behavior influence participants’ behaviors or responses. *Participant reactivity effects* are behaviors and responses made by participants that are due to the study setting or attributes about the experimenter from their mere presence to more specific aspects such as their age, gender, or race. *Observer effects* are the biased observations made by experimenters of participants’ behaviors or responses (Rosenthal & Rosnow, 2009).

Expectancy Effects

Expectancy effects occur when the experimenter’s a priori expectations regarding participants’ behavior change how experimenters interact with participants and alter participants’ behavior to conform to experimenters’ beliefs. While expectancy effects are rarely as blatant as Clever Hans, they are frequently difficult to avoid and can produce sizable shifts in experimental outcomes.

In a foundational study on expectancy effects, Rosenthal and Fode (1963) recruited a group of students to be “experimenters” in a study. Their job was to show photographs of individuals to research participants and ask them to judge the probability that the pictured individuals were successful in life. Half of the student experimenters were told that the people in the photos had previously been rated as successful, while the other half were told those in the photos were previously rated as failures. The experimenters were told that they would be paid more if they did a “good job” (i.e., replicated previous findings). However, the experimenters were also instructed to limit their interactions with participants and to only read the standardized instructions. Despite limiting interactions and reading the standardized instructions, participants’ ratings of the photos were nonetheless consistent with the expectations given to the experimenters (Rosenthal & Fode, 1963). These results provide a salient example of how researchers alter their behavior based upon prior beliefs, causing participants to respond in a way that matches researcher expectations.

In a review, Atwood and colleagues (2020) examined the extent of such expectation effects. Specifically, they examined the effect of synchronous interpersonal movement on increased prosocial behavior (i.e., “mirroring” another person – if your conversation partner puts their hand to their chin, you also put your hand to your chin). They found that when experimenters had an idea of how the participants *should* act based on the condition they were assigned, participants were more likely to move in a synchronous way that was related to the participants’ subsequent prosocial behaviors during the experiment. Furthermore, these effects are moderated by experimenter expectancy (Rennung & Göritz, 2016). Specifically, experimenters’ expectations and subsequent movements could explain why participants exhibited increased prosocial behaviors in the experiments. These findings suggest that the experimenter’s knowledge of the study’s purpose or motivation for certain outcomes can alter participant behaviors, even when these behaviors are thought to be well controlled.

A specific type of expectancy effect, the *Pygmalion effect*, occurs when experimenters expect to observe increased performance for a particular group. The Pygmalion effect has most prominently been studied in academic settings. Rosenthal and Jacobson (1968) conducted a study in which elementary school teachers were told that a certain group of students in their classrooms would be "intellectual bloomers" and outperform their peers by the end of the school year. In reality, those students were chosen at random. The students who were labeled intellectual bloomers had significantly greater improvements in their IQ scores throughout the school year than the other students. A follow-up study revealed that teachers who formed more positive attitudes and expectancies of certain students showed increased attention and support, offered more challenging learning materials, interacted more, and gave more feedback to those students (Brophy & Good, 1970). These findings are analogous to a *self-fulfilling prophecy*; when people have predictions or beliefs of a certain outcome, their resulting behaviors align to fulfill that belief (Rosenthal, 1973).

In some cases, experimenter behavior (or study design more broadly) can alter participant behavior indirectly by giving them cues to an experiment's purpose or hypothesis, or, more generally, how they are expected to behave. These are referred to as *demand characteristics* (Orne, 1962), which can be due to direct communication or unintentional hints from the experimenter or other participants (Klein et al., 2012). Participants typically act in one of three ways in response to the demands placed on them in an experiment: they may exhibit behaviors to support the hypothesis (*good-subject effect*; Nichols & Maner, 2008), exhibit behaviors to disprove the hypothesis (*bad-subject effect*; Argyris, 1968), or they can ignore the demands and negate their potential effects. In one study, participants were told by a confederate (i.e., a researcher acting as a participant) that, despite what they would be told in the experiment, the "true" purpose of the study was to test whether people prefer things presented in their left or right visual field and that the left side should be preferred (the actual tendency without manipulations would be to prefer things on the right). The researchers found that students acted in a manner that was consistent with the provided expectation and preferred stimuli presented on the left side (i.e., the good-subject effect; Nichols & Maner, 2008).

While it may be safe to assume that, in general, researchers would not purposely inform participants of the hypothesis, this example shows that demand characteristics and the participant's expectations of the experiment's purpose can influence the results. Thus, even implicit hints made by the experimenter could elicit similar effects. Additionally, participant crosstalk may contribute to the spread of any purported hypotheses in the study (Edlund et al., 2009). This is particularly relevant to studies that involve sampling from a restricted pool of participants (e.g., students from a particular college course; see also Edlund et al., 2017).

Another related phenomenon occurs when participants develop beliefs about a study's purpose after they have learned about other conditions in the experiment. Participants act as good subjects to the extent their beliefs match the actual hypotheses. In particular, research shows that, if participants perceive their assigned condition to be at a disadvantage compared to other conditions in the experiment (e.g., they have been assigned to a control or placebo condition), they may try harder to compensate for the disadvantage. This behavioral pattern is sometimes called the *John Henry effect*. In a seminal example, researchers examined the effectiveness of "performance contractors," or teachers who were paid according to how well their students performed. The performance contractors were compared to a control group of teachers who taught the standard curriculum in their classrooms. Initially, the results showed that there were no differences in the students' reading and math performance between the classrooms with performance contractors and control teachers. However, closer examination showed that the students in the control classrooms performed much higher than the typical classroom during previous years, suggesting that the control teachers actively worked harder to increase their students' scores to overcome a perceived disadvantage of being compared to the performance contractors (Saretsky, 1972).

While the John Henry effect may be considered a participant reactivity effect, the type of control group used or how expectations of the control group are masked are important choices for experimenters to consider during the experimental design step. In particular, a control manipulation (or lack of) that is too obvious may exacerbate the effect. For example, a participant recruited for an experiment testing the effectiveness of a behavioral intervention may intuit they are in a control group if they do not receive an intervention, and their performance on the post-intervention assessments may change. Knowledge of being in a control group may motivate participants to try harder (i.e., the John Henry effect) or may have the opposite effect and cause participants to lose interest and stop trying (i.e., bad-subject effect). Both outcomes could lead to incorrect conclusions.

Participant Reactivity Effects

While researchers typically attempt to experimentally control factors that may affect participants, the participant may inevitably respond to aspects not intended to influence their responses. Critically, many of these aspects are also not in the experimenter's control. These range from global characteristics (e.g., their simple presence) to more individual aspects (e.g., experimenter's age, gender, or race).

For example, the *Hawthorne effect* – named after a series of experiments conducted in an electrical plant near Chicago called Hawthorne Works in the early 1900s – occurs when participants are simply aware of being in a study and/or being observed by experimenters. The experimenters wanted to test whether different factors (e.g., pay, light levels, and breaks) increased worker productivity. However, regardless of the factor they manipulated, worker performance increased while experimenters were present and returned to baseline levels after experiments ended. The researchers concluded that the performance increases were simply due to the workers' awareness of being studied (French, 1953). Similar effects have been demonstrated across many areas of research (McCambridge et al., 2014). For example, interviews before an election led to an increased probability of voting compared to interviews after an election (Granberg & Holmberg, 1992). Similarly, completing initial alcohol-use or smoking questionnaires led to lower self-reported drinking-related problems or smoking on a later questionnaire compared to those who did not have initial questionnaires (McCambrdige & Day, 2007; Murray et al, 1988). The experimenter's decisions while designing the procedures of the experimenter, such as when or how often to implement questionnaires, play a vital role in how these effects are manifested and ultimately affect the outcomes of the study.

In addition to effects that occur simply due to an experimenter's presence, more specific changes in participant behavior may occur as a function of more specific attributes of the experimenters (e.g., gender, race, beliefs, or personality). These effects are largely rooted in the effects of *stereotypes* – beliefs about characteristics, attributes, and behaviors of certain groups (see Hilton & von Hippel, 1996 for a review). Participants' stereotypes about experimenters can influence behaviors and responses, thus impacting study outcomes. Indeed, participants are more likely to exhibit a good-subject effect when they have a more positive attitude towards the experimenter (Nichols & Maner, 2008).

For instance, some of these stereotypes are rooted in the experimenter's race or gender. Nichols and Maner (2008) found that, when participants interacted with an opposite-sex experimenter, single participants were more likely to behave as a good subject. Furthermore, male participants tend to report lower pain in the presence of a female experimenter than a male experimenter (Asklaksen et al., 2007; Levine & De Simone, 1991). Similarly, female participants reported a more positive attitude toward casual sex when they completed a survey with a female experimenter than a male experimenter (McCallum & Peterson, 2015). Regarding race effects, Black participants performed worse on a verbal task than White participants when the experimenter was White compared to other races (Marx & Goff, 2005). Experimenters' race and gender can also affect participants' physiological responses (e.g., heart rate; Thorson et al., 2019).

Besides an experimenter's biosocial attributes, an experimenter's attire may influence how participants behave. In Morocco (a predominantly Muslim country), the experimenter's religious dress affected participants' responses to religiously sensitive questions. Participants were significantly less likely to give highly religious answers when interviewed by a secular male compared to a Muslim woman wearing a hijab. These findings suggest that participants' responses to religiously sensitive questions can be influenced by *social desirability* – the tendency to respond in a more socially acceptable way than would be their "true" answer (Benstead, 2014). However, what is deemed socially acceptable for one participant may be the opposite for another or for the experimenter. One study found that increased social desirability was associated with a *lower* likelihood of being a good subject. This may be because the idea of what is socially desirable may differ between participants and experimenters (Nichols & Maner, 2008).

Not only can the physical attributes of experimenters affect participants' behaviors, but their personality attributes (e.g., a need for approval, hostility, warmth, or authoritarianism) may also affect behavior (Rosenthal, 1963; Rosenthal & Rosnow, 2009). Having a warm manner involves verbal and non-verbal cues (e.g., taking time for introductions or making eye contact) and stands in contrast to having a cold manner, wherein one lacks emotion and does not engage in conversations. These characteristics affected participants' levels of disclosure in a study involving writing about a traumatic event (Rogers et al., 2007). When experimenters exhibited a warm manner, participants disclosed more information compared to experimenters with monotone voices and minimal eye contact. Studies have also illustrated the effects of experimenters' confidence levels on outcomes; higher levels of confidence in experimenters caused significantly smaller allergic reactions in participants given an inert cream following an allergy skin prick test (Howe et al., 2017). Such work demonstrates that biosocial and psychosocial characteristics that may be out of the experimenter's conscious control can play a role in participant responses and the outcomes of the experiment.

The level of the experimenter's status, as signified by professionalism and/or social status, can also affect participant behaviors. Professionalism can be described as perceptions of a person's competence or skill; this can be indicated in a number of ways, such as a person's education level (e.g., professor vs. student) or the way one dresses (e.g., lab coat vs. a tank top). For example, higher levels of experimenter professionalism have been linked to increased pain tolerance in keeping one's hand in a bath of cold water (a typical method of testing pain tolerance) for both male and female participants (Kállai, et al., 2004). Additionally, participants who were tested on a pressure pain threshold task by a professor reported higher pain thresholds than participants who were tested by a student (Modic-Stanke & Ivanec, 2016).

These status-related effects are similar to *obedience effects* where participants' behaviors change in response to an authority figure instructing them to behave in some way. For example, in the classic Milgram experiment, a study falsely advertised as a study on the effect of punishment on memory-participants were assigned to be a "teacher" along with a confederate researcher assigned to be a "learner." The teacher was to administer an electric shock, with increasing intensity, for each incorrect answer from the learner. When the teacher wanted to stop the experiment, the experimenter asked them to continue, stating: "You have no other choice; you must go on." The results showed that all participants administered high levels of electrical shocks, and 65% administered the maximum level (despite the learners frequently exhibiting signs of distress when doing so; Milgram, 1963). Follow-up studies to the Milgram paradigm found that some properties of the experimenter or experimental setting led to stronger obedience effects, including the status of the experimenter or prestige of the institution (e.g., 47.5% obedience rate at Bridgeport vs. 62.5% at Yale; Haslam et al., 2014). In summary, the studies on experimenter attribute effects appear to be widespread and depend on a number of experimenter characteristics.

Observer Effects

Some research may not require experimenters to directly interact with participants. However, experimenter bias can still alter the observed results. For instance, this can happen when experimenters are tasked with recording observations of participants (e.g., watching a recording of children playing and noting whenever children display aggressive behaviors). Indeed, such observations can be influenced by observer beliefs, particularly when subjective measurements are made (Hoyt, 2000). Like expectancy effects, researchers who are aware of participants' conditions may be biased in their judgments of the participants' responses (*observer effect*), which may be driven by the researchers' confirmation bias. Tuyttens and colleagues (2014) conducted several experiments to examine this possibility. In one example, they asked veterinary students to observe and record positive and negative social interactions of pigs. They were told that one video showed pigs who were selected for high social breeding value and the second showed a control group of pigs. In reality, they were the same video, but one was altered to be unrecognizable as the same video (mirror reversed, brightness changed, etc.). The students scored the ratio of positive to negative interactions higher for the pigs that were believed to have higher social breeding value than control pigs.

A broader example of how our beliefs can impact our observations is seen in the common notion among parents that sugar induces hyperactive behaviors in their children. Parents were asked to rate their children's behaviors on two separate occasions; on one occasion, they were told that their children had received a drink containing sugar, and on the other, they were told their children had received a drink not containing sugar. While the children received a sugary drink on both occasions, parents reported that their children were more restless, impulsive, fidgety, and distracted after drinking the drink believed to contain sugar (Spring & Alexander, 1989).

In general, numerous studies across different research domains illustrate how experimenters' beliefs can bias their observations and thus affect the outcomes of the study.

Experimenter Effects During Data Analyses and Interpretation

After data have been collected, experimenter effects can influence how the data are analyzed and interpreted. Typically, data sets can be analyzed in several ways, and the decisions that experimenters make during these analyses affect both the statistical values obtained and the conclusions drawn from the analyses. Motivations for selecting particular analyses may be rooted in confirmation bias (e.g., selecting only the data that would support one's hypothesis). Another motivation may be due to *publication bias* – the tendency for peer-review journals to publish studies with significant results more than non-significant results. Thus, experimenters may feel compelled to run analyses until they obtain significant and, thus, publishable results.

One method of doing so, called *p-hacking*, refers to practices of selecting or manipulating data to better support one's hypothesis, such as obtaining a significant p-value or a larger effect size (Head et al., 2015). Examples of these practices include conducting analyses partway through data collection to decide whether to continue collecting data (e.g., researchers would stop collecting data if they found a significant result), recording many variables and deciding which to report only after analyzing the data (e.g., "cherry-picking" only the variables with significant results) or choosing to include or exclude certain participants and covariates in the analyses (e.g., removing data points that would disconfirm the hypothesis; John et al., 2012; Simmons et al., 2011). This process may consist of repeated analyses until there is a significant result. The practice of *p*-hacking can result in biased interpretations of the data and increases the likelihood of false-positive results, where researchers conclude that an effect exists based upon spurious evidence (Friese & Frankenbach, 2020). An influx of false positives in a literature can drastically impact the overall consensus in a particular research area. Indeed, the estimated mean effect size found in meta-analyses is likely inflated by *p*-hacking techniques (Head et al., 2015).

After the data are analyzed, another bias that may arise is known as *HARKing* – hypothesizing after results are known (Kerr, 1998). The scientific method entails that experimenters state their hypotheses *before* both data collection and statistical analyses. However, this requirement is often violated, and it can be difficult for readers of a published study to ascertain when hypotheses were truly formed. Here, another common and well-known human bias can be at play – *hindsight bias* – the tendency to conclude that an event was predictable only after it has occurred. A classic study of hindsight bias began in 1972, before President Nixon's trips to Peking, China, and Moscow, Russia. Participants were first asked to judge the likelihood that certain events would occur during those trips, such as President Nixon meeting with Chinese Chairman Mao Zedong or the USA and USSR agreeing to a joint Space program. After the president returned, the participants were then asked to remember their original predictions as well as whether they believed the event occurred. The remembered predictions were generally higher for events that occurred or were believed to occur than events that did not occur, and participants seldom perceived that they were surprised by what had actually happened, despite their original predictions (Fischhoff & Beyth, 1975). Because we are unable to reconstruct the feeling of uncertainty that preceded an event, when we think back, we often feel like we "knew it all along." Similarly, experimenters' hindsight bias may play a role in how they interpret and report their findings after the experiment (Munafò et al., 2017).

Hindsight bias combined with a desire to form a coherent or more publishable narrative of the data, due to publication bias, may drive researchers to alter their hypotheses once they see what the data actually show. For example, if a researcher finds a significant relationship or effect that was not originally predicted, it may be easy to add this prediction to their original hypotheses to better fit the data or remove hypotheses that were not statistically supported. Not only are *p*-hacking and HARKing poor scientific practices, but they also threaten the reproducibility of studies; "significant effects" found via *p*-hacking are more likely to be false positives and less likely to be replicated (Munafò et al., 2017). The replicability of experimental results is an essential part of the scientific method and allows future researchers to support those results and formulate subsequent theories based on those results. Thus, the inability to replicate studies has potentially grave consequences for areas of research where theories are grounded on unreproducible experimental work. For example, if a policy about an intervention for a mental health disorder was based on conclusions from a *p*-hacked result, it would hurt both the scientific community and the clinical population affected by the intervention.

We note that, while HARKing and *p*-hacking are significant problems, this does not negate the importance or legitimacy of purposely exploratory research in which no specific hypotheses are stated. For both exploratory work and hypothesis-based research, it is important for experimenters to state the aims of their research from the start and to report which observations were made after the fact. Further, making a point to replicate studies can aid addressing biases in data analyses and interpretation.

In addition to HARKing and *p*-hacking, when results are consistent with the experimenter's biases, there is also the tendency of confirming a hypothesis based on "good enough" evidence rather than looking for alternative explanations (Bishop, 2020). When a reliable effect is observed, researchers often see this as evidence for their original or preferred hypothesis without considering how alternative explanations may be supported by the same evidence or controlling for these potential confounds. This bias for accepting a convenient explanation can mask better explanations of an effect. For example, before the use of a control group became the "gold standard" of evaluating the effectiveness of a medical treatment (Vallier & Timmerman, 2008), many studies concluded that if the drug improved patients' symptoms, it was effective. Though methodological standards have since improved, this example illustrates that, without the use of a control group, experimenters may incorrectly conclude that a drug is effective without considering alternative explanations (e.g., potential placebo effects or the treatment group improving naturally over time as much as a control group would have; Bishop, 2020).

In a more recent example, it was hypothesized that dyslexia has a neurological basis, and a relationship between dyslexia and atypical brain responses in response to speech was observed. The favored explanation was that this was due to atypical brain organization in the language area associated with dyslexia (Shaywitz et al., 2006). However, an alternative explanation would be that atypical brain responses to speech are a *consequence* of being a poor reader. A later study found that adults who had never been taught to read also had atypical brain organization for spoken language, thus providing evidence against the original explanation that atypical brain responses lead to dyslexia (Dehaene & Cohen, 2011). It is important to note, however, that a search for alternative explanations is not the same as HARKing since HARKing presents explanations generated after testing the original hypotheses, whereas a study that adds alternative explanations of the results in the discussion includes both the original hypotheses as well as future hypotheses to test.

Finally, in the last step of writing up a study manuscript, one potential experimenter effect may manifest in *citation bias* or the tendency to not cite literature contrary to one's views. As we have already seen, we tend to reconstruct our memories over time to become more in-line with our pre-existing representations of information (Bartlett, 1932; Bishop, 2020). We are also more likely to ignore evidence inconsistent with our beliefs (Vicente & Brewer, 1993). In fact, when we encounter contradictory evidence, we may experience *cognitive dissonance*-like effects – a state of psychological stress that occurs when we encounter outcomes that counter our beliefs, ideas, or values. Because of this, we may downplay or distort contradictory evidence to alleviate those feelings (Duyx et al., 2017).

Consistent with this, one review found that researchers examining possible links between violent video game play and increases in aggression frequently highlight previous work that has shown such a relationship without mentioning contradictory evidence (Ferguson, 2015). Similarly, a review of the depression intervention literature reported both a bias against publishing null findings (i.e., those that found that the interventions used were not effective at improving some depression outcome) as well as a bias against citing published null findings (De Vries et al., 2018). Additionally, referenced null findings were often spun to give a more positive impression of the intervention. These examples suggest that researchers may be biased toward which previous studies they choose to discuss or how they want to spin the story. As such, this "cherry-picking" of previous literature can further support the experimenters' results and intended take-home message from their study.

Recommendations for Minimizing Experimenter Effects

Minimizing Effects Before the Experiment

Experimenters can take precautions throughout a study to minimize experimenter effects. When conducting a literature review, the search should be thorough and aim to include multiple perspectives on a given topic. Beyond just citing individual studies, it is important to identify meta-analyses or systematic reviews that take into consideration as many related results as possible, including those that are unpublished (i.e., "gray literature"). Because we are prone to selective exposure, this may be easier said than done. Creating an objective list of search terms or asking another collaborator or co-author to also conduct a literature search may help reduce this bias (Winchester & Salji, 2016).

Minimizing Effects During the Experiment

While designing the study, it is important to consider a number of explicit procedures in the experiment to reduce unwanted biases during data collection. As we saw earlier, expectancy effects are due to the experimenter being aware of the conditions of the participants, which may result in experimenter behaviors toward participants that differ by condition and, in turn, cause the participant to behave in accordance with the experimenter's expectations of those conditions. In the two examples of experimenter effects on ratings of photos and on prosocial behavior, the experimenter unintentionally provided positive feedback, such as facial expressions or

https://doi.org/10.1017/9781009010054.012 Published online by Cambridge University Press

body language, when the participants' responses or behaviors were in line with the experimenter's hypothesis. Similarly, in the case of Pygmalion effects in the classroom, teacher expectations of which students should have better outcomes led to behaviors to support those expectations, such as giving the students more attention or providing more feedback. Overall, the impact of an experimenter with knowledge of conditions has been shown to have a large effect on participant behaviors. In a meta-analysis, studies with aware experimenters reported more statistically significant results and larger effect sizes than studies with unaware experimenters (Holman et al., 2015).

The best way to minimize these effects is by masking any information about the conditions from the experimenter throughout a study. The gold-standard design is one in which both the participants and experimenters are unaware of which conditions the participants are assigned. We note that this has historically been termed a "double-blind" study. However, because this type of study does not literally involve making the participants or experimenter unable to see (and thus could be considered ableist), below we use terms such as "masking" or "unaware," which are more accurate descriptions of the intention of experimental procedures (see Morris et al., 2007).

Reducing participant awareness of conditions can minimize experimenter effects in two ways. First, keeping participants unaware to their conditions can help prevent the John Henry effect, as participants would be less likely to intuit the condition they are assigned. Second, if the experimenter does not know which group the participant belongs to, they should have no expectancy of how the participant should behave. This can reduce systematic acts, such as facial expressions, body language, and tone of voice, made by experimenters towards the participant, as well as observations made about participants' behaviors or responses, based on their assigned condition. Rosenthal and Rosnow (2008) outlined some strategies to help maintain experimenter masking. For example, increasing the number of experimenters can lead to fewer instances of experimenters learning of conditions during data collection and would randomize expectations of experimenters when they do form expectancies. Though increasing the number of experimenters might also increase the likelihood of procedural errors, one can also continuously monitor experimenters to ensure that they consistently follow a standard protocol across all participants and conditions and identify biases when they occur. In the case of making observations, having more than one experimenter and checking for inter-rater reliability would decrease the likelihood of experimenter effects. To determine whether both experimenters and participants were unaware of conditions, it may be useful to include a post-study questionnaire to probe whether they were aware of the hypotheses (see Chapter 12 in this volume).

It is also important to consider whether and how the experimenters' physical, biosocial, or psychosocial characteristics (e.g., attire, age, gender, or personality) affect the participant. To eliminate experimenter effects, one option is to remove experimenter contact with participants when possible (e.g., having written instructions; Rosenthal & Rosnow, 2008). However, in many cases, an experimenter is necessary. Regarding experimenter attire, requiring all experimenters to wear a lab coat may minimize attire-driven demand characteristics (Nichols & Edlund, 2015). To control for between-experimenter differences, one could use the same researcher throughout the entire study. While using the same experimenter for the entirety of the experiment can help reduce non-interactional effects as the experimenters' attributes are held constant for all participants, this would likely introduce other issues such as the possibility of the experimenter becoming aware of conditions and the infeasibility of time commitment. As discussed more in the next section, an alternative option is to estimate or account for various experimenters' attributes on subsequent participant behaviors in statistical analyses, rather than attempting to minimize or dismiss their effects altogether.

Some of these reactivity and expectancy effects during data collection may also be reduced by conducting experiments online where experimenters interact with participants via online chat or video conference. In online studies, participants may receive fewer body language cues from the experimenter, reducing the chance that these cues have an influence. However, online research introduces other potential issues related to the internal validity of the experiment, such as not being able to control the participant's environment, so it is important to weigh the benefits of online research with these costs (Nichols & Edlund, 2015).

Minimizing Effects After the Experiment

Once the study is designed and hypotheses are formulated, an informal way of evaluating potential biases, arising either during data collection or in the analyses, is to present the study to other researchers in the field, especially those that may have different views or hypotheses, to discuss appropriate statistical approaches. A more formal way can be done through preregistered reports. *Preregistration* is a process that involves publicly sharing one's a priori hypotheses and planned statistical analyses prior to data collection. Preregistration may include information on the number of participants, treatment of outliers, and which comparisons will be made with specific statistical analyses. Preregistration helps in assessing experimenter bias at the initial stages of designing a study and prevents bias during data analyses (e.g., *p*-hacking and HARKing), as the experimenters' hypotheses and statistical methods are stated beforehand.

Recently, there has been a push for preregistering studies, particularly in social and behavioral sciences (Nosek et al., 2017). Websites, such as the Open Science Framework (osf.io) and ClinicalTrials.gov, provide a place for researchers to register their studies and publicly share their data or other relevant materials. Preregistering a study and publicly sharing collected data after the study increases the scientific community's trust that the results of the study are not merely due to interpretation effects. The Registered Reports initiative also promotes good science practices and may help reduce the biased interpretation of results. In a two-step process, researchers submit their detailed study rationale, experimental protocol, and statistical analysis plan to a peer-reviewed journal. After being approved by reviewers, the journal offers tentative acceptance of the study provided that the authors follow their stated plans. This allows researchers to make conclusions about their data without the pressure of publication bias toward finding significant effects (Nosek & Lakens, 2014). While preregistration can reduce some experimenter biases, one downside is that stating these hypotheses and statistical analyses early in the process may discourage post-hoc exploratory analyses. However, this can be addressed by clearly stating which analyses are preregistered and which analyses are post-hoc or exploratory, thus enabling readers to judge the results for themselves.

Measuring or Accounting for Experimenter Effects

Experimenter effects can have implications in how the results of a study will be replicated and generalized. For example, if the results of a study are due to particular attributes of specific experimenters, they may not be replicable with other experimenters. While researchers can do their best to minimize experimenter effects across the duration of a study, it may not be feasible (or practical) to eliminate them. Because of this, possibly a more effective approach is to actually estimate their influence on participant responses to better understand their role in a study. To estimate the expectancy effect, Rosenthal and Rosnow (2008) suggested employing an expectancy control group design. In this design, the experimenter can compare the effect of the experimenter's expectancy against the effect of the true manipulation in the experiment.

For example, a study could include one condition in which participants receive an experimental treatment and one condition in which participants receive a control treatment. Half of the researchers would be told the true conditions (experimental group received the experimental treatment and control group received the control treatment), while the other half would be told the opposite (experimental group received the control treatment and control group received the experimental treatment). Statistical analyses would then examine the effects of both the manipulated treatment and experimenter expectancies on the outcome measures (Rosenthal & Rosnow, 2008). Two variables are manipulated in this kind of design: the treatment variable and the experimenter's expectation. Because of this, the number of participants per group would have to be doubled to maintain the same level of statistical power to examine both effects. However, a design like this provides insight on the magnitude of the effect of experimenters' expectancies on participants' responses and behaviors.

Rather than measuring experimenter effects directly by creating additional conditions within an experiment, researchers can also estimate their effects by including them as a variable in analyses. While researchers generally agree that any variables outside of the true manipulation of the experiment should be controlled or accounted for, it is common to overlook or ignore the impact of experimenters themselves. The implications of this may lead to an overgeneralization of the results. One cannot rule out that conclusions drawn from the results are not unique to the experimenters included in the experiment. As discussed throughout this chapter, there are abundant possibilities in which experimenters affect participants' behaviors and study outcomes, and these should be considered when modeling the data. One approach to analyzing data from an experiment may be to calculate the effect of the independent or predictor variables on the outcome variables in linear regression. Under such an approach, we can consider the experimenters themselves as another predictor variable and expand the model to account for these factors (e.g., the experimenters' genders, races, and ages). By statistically controlling for the attributes, we can estimate their impact on the study results as well as examine whether inclusion of those variables substantially changed the results (Yarkoni, 2020). However, it is difficult to disentangle where experimenter effects come from. For example, expectancy effects may interact with participant reactivity effects and may not be readily teased apart. Furthermore, many experiments might be underpowered to detect such effects.

Are Experimenter Effects Always a Bad Thing?

The preceding sections describe cases in which experimenters' biases, whether intentional or not, can affect the results and conclusions of a study. In many cases, this is something that researchers want to minimize as much as possible, as it may confound the true effect of the variables being tested. However, experimenter effects may be useful in maximizing the impact of a particular intervention or phenomenon being studied to positively alter participant behaviors. For example, in a clinical psychology experiment examining the efficacy of a behavioral intervention (e.g., an intervention to reduce anxiety), rather than trying to minimize the possible impact of an experimenter's interactions with participants, it may be more clinically valuable to harness the effects of those attributes that can positively impact the intervention (e.g., experimenters' warmth and openness towards participants; Rogers et al., 2007). Another example may be in taking advantage of the Pygmalion effect in the classroom or workplace, which would be useful both for experimental purposes and as general recommendations for creating better school and work environments. Raising a manager's expectations of their employees may subsequently lead to increasing employees' motivation and produce better productivity (Eden, 1984). Similarly, as previous evidence has shown that teachers' expectations of their students can significantly impact their educational outcomes, one could also harness these effects for improving student outcomes (Weinstein, 2018).

Conclusion

In summary, experimenter effects can occur in virtually every step of research from literature review to write-up of manuscripts for publication. Many steps can be taken to minimize these effects, but likely some experimenter effects will remain even with the strictest protocols; sometimes, researchers may want to take advantage of them rather than minimizing them. Thus, using approaches to estimate experimenter effects may be particularly useful in understanding their true impact on outcomes.

References

Argyris, C. (1968). Some unintended consequences of rigorous research. *Psychological Bulletin*, *70*(3, Pt.1), 185–197. https://doi.org/10.1037/h0026145

Asklaksen, P. M., Myrbakk, I. N., Hoifodt, R., S., & Flaten, M. A. (2007). The effect of experimenter gender on autonomic and subjective responses to pain stimuli. *Pain*, *129*(3), 260–268.

Atwood, S., Mehr, S. A., & Schachner, A. (2020). Expectancy effects threaten the inferential validity of synchrony-prosociality research [Preprint]. *PsyArXiv*. https://doi.org/10.31234/osf.io/zjy8u

Bartlett, F. C. (1932). *Remembering: A Study in Experimental and Social Psychology*. Cambridge University Press.

Benstead, L. J. (2014). Does interviewer religious dress affect survey responses? Evidence from Morocco. *Politics and Religion*, *7*(4), 734–760. https://doi.org/10.1017/S1755048314000455

Bishop, D. V. M. (2020). The psychology of experimental psychologists: Overcoming cognitive constraints to improve research. *Quarterly Journal of Experimental Psychology*, *73*(1), 1–19. https://doi.org/10.1177/1747021819886519

Brophy, J. E. & Good, T. L. (1970). Teachers' communication of differential expectations for children's classroom performance: Some behavioral data. *Journal of Educational Psychology*, *61*(5), 365–374. https://doi.org/10.1037/h0029908

Dehaene, S. & Cohen, L. (2011). The unique role of the visual word form area in reading. *Trends in Cognitive Sciences*, *15*(6), 254–262. https://doi.org/10.1016/j.tics.2011.04.003

De Vries, Y. A., Roest, A. M., de Jonge, P., et al. (2018). The cumulative effect of reporting and citation biases on the apparent efficacy of treatments: The case of depression. *Psychological Medicine*, *48*, 2453–2455. doi:10.1017/S0033291718001873

Duyx, B., Urlings, M. J. E., Swaen, G. M. H., Bouter, L. M., & Zeegers, M. P. (2017). Scientific citations favor positive results: A systematic review and meta-analysis. *Journal of Clinical Epidemiology*, *88*, 92–101. https://doi.org/10.1016/j.jclinepi.2017.06.002

Eden, D. (1984). Self-fulfilling prophecy as a management tool: Harnessing Pygmalion. *The Academy of Management Review* 9(1), 64.

Edlund, J. E., Cuccolo, K., Irgens, M. S., Wagge, J. R., & Zlokovich, M. S. (2021). Saving science through replication studies. *Perspectives on Psychological Science*, *17*(1), 216–225. https://doi.org/10.1177/1745691620984385

Edlund, J. E., Lange, K. M. Sevene, A. M., et al. (2017). Participant crosstalk: Issues when using the Mechanical Turk. *The Quantitative Methods in Psychology*, *13* (3), 174–182.

Edlund, J.E., Sagarin, B.J, Skowronski, J.J., Johnson, S., & Kutter, J. (2009). Whatever happens in the laboratory stays in the laboratory: The prevalence and prevention of participant crosstalk. *Personality and Social Psychology Bulletin*, *35*, 635–642.

Ferguson, C. J. (2015). Pay no attention to that data behind the curtain: On angry birds, happy children, scholarly squabbles, publication bias, and why betas rule metas. *Perspectives on Psychological Science*, *10*(5), 683–691. https://doi.org/10.1177/1745691615593353

Fischhoff, B. & Beyth, R. (1975). "I knew it would happen": Remembered probabilities of once-future things. *Organizational Behavior & Human Performance*, *13*(1), 1–16. https://doi.org/10.1016/0030-5073(75)90002-1

Forster, K. L. (2000). The potential for experimenter bias effects in word recognition experiments. *Memory and Cognition*, *28*, 1109–1115.

French, J. R. P. (1953), Experiments in field settings. In L. Festinger & D. Katz (eds.), *Research Methods in the Behavioral Sciences* (pp. 98–135), Holt, Rinehart and Winston.

Friese, M. & Frankenbach, J. (2020). *p*-Hacking and publication bias interact to distort meta-analytic effect size estimates. *Psychological Methods*, *25*(4), 456–471.

Granberg, D. & Holmberg, S. (1992). The Hawthorne effect in election studies: The impact of survey participation on voting. *British Journal of Political Science*, *22*(2), 240–247.

Hart, W., Albarracín, D., Eagly, A. H., et al. (2009). Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psychological Bulletin*, *135*(4), 555–588.

Haslam, N., Loughnan, S., & Perry, G. (2014). Meta-Milgram: An empirical synthesis of the obedience experiments. *PloS One*, *9*(4), e93927. https://doi.org/10.1371/journal.pone.0093927

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of *p*-hacking in science. *PLoS Biology*, 13, e1002106.

Hilton, J. L. & von Hippel, W. (1996). Stereotypes. *Annual Review of Psychology*, *47*(1), 237. https://doi.org/10.1146/annurev.psych.47.1.237

Holman, L., Head, M. L., Lanfear, R., & Jennions, M. D. (2015). Evidence of experimental bias in the life sciences: Why we need blind data recording. *PLoS Biology*, *13*(7). https://doi.org/10.1371/journal.pbio.1002190

Howe, L. C., Goyer, J. P., & Crum, A. J. (2017). Harnessing the placebo effect: Exploring the influence of physician characteristics on placebo response. *Health Psychology*, *36*(11), 1074–1082. https://doi.org/10.1037/hea0000499

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods*, *5*, 64–86. doi:10.1037//1082-9S9X.5.1.64

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Kállai, I., Barke, A., & Voss, U. (2004). The effects of experimenter characteristics on pain reports in women and men. *Pain*, *112*(1), 142–147. https://doi.org/10.1016/j.pain.2004.08.008

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2, 196–217.

Klecka, H., Johnston, I., Bowman, N. D., & Green, C. S. (2021). Researchers' commercial video game knowledge associated with differences in beliefs about the impact of gaming on human behavior. *Entertainment Computing*, *38*, 100406. https://doi.org/10.1016/j.entcom.2021.100406

Klein, O., Doyen, S., Leys, C., et al. (2012). Low hopes, high expectations: Expectancy effects and the replicability of behavioral experiments. *Perspectives on Psychological Science*, *7*, 572–584.

Levine, F. M. & De Simone, L. L. (1991). The effects of experimenter gender on pain report in male and female subjects. *Pain*, *44*, 69–72.

Marx, D. M. & Goff, P. A. (2005). Clearing the air: The effect of experimenter race on target's test performance and subjective experience. *British Journal of Social Psychology*, *44*, 645–657.

McCallum, E. B. & Peterson, Z. D. (2015). Effects of experimenter contact, setting, inquiry mode, and race on women's self-report of sexual attitudes and behaviors: An experimental study. *Archives of Sexual Behavior*, *44*, 2287–2297.

McCambridge, J. & Day, M. (2007). Randomized controlled trial of the effects of completing the Alcohol Use Disorders Identification Test questionnaire on self-reported hazardous drinking. *Addiction*, *103*, 241–248

McCambridge, J., Witton, J., & Elbourne, D. R. (2014). Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects. *Journal of Clinical Epidemiology*, *67*(3), 267–277. https://doi.org/10.1016/j.jclinepi.2013.08.015

Meier, A., Domahidi, E., & Günther, E. (2020). *Computer-Mediated Communication and Mental Health: A Computational Scoping Review of an Interdisciplinary Field*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190932596.013.4.

Milgram, S. (1963). Behavioral study of obedience. *The Journal of Abnormal and Social Psychology*, *67*(4), 371–378. https://doi.org/10.1037/h0040525

Modic-Stanke, K. & Ivanec, D. (2016). Pain threshold: Measure of pain sensitivity or social behavior? *Psihologija*, *49*(1), 37–50. https://doi.org/10.2298/PSI1601037 M

Morris, D., Fraser, S., & Wormald, R. (2007). Masking is better than blinding. *BMJ: British Medical Journal (International Edition)*, *334*(7597), 799.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., et al. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, 0021. https://doi.org/10.1038/s41562-016-0021

Murray M., Swan A. V., Kiryluk S., & Clarke, G. C. (1988). The Hawthorne effect in the measurement of adolescent smoking. *Journal of Epidemiology & Community Health*, *142*, 304–306.

Nichols, A. L. & Edlund, J. E. (2015). Practicing what we preach (and sometimes study): Methodological issues in experimental laboratory research. *Review of General Psychology*, *19*(2), 191–202.

Nichols, A. L. & Maner, J. K. (2008). The good-subject effect: Investigating participant demand characteristics. *Journal of General Psychology*, *135*(2), 151–165.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175–220. doi:10.1037/1089-2680.2.2.175

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2017). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Nosek, B. A. & Lakens, D. (2014). Registered Reports: a method to increase the credibility of published results. *Journal of Social Psychology*, *45*, 137–141.

Orne, M.T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, *17*, 776–783.

Pfungst, O. (1911). *Clever Hans (The Horse of Mr. von Osten*). Holt, Rinehart, & Winston,.

Rennung, M. & Göritz, A. S. (2016). Prosocial consequences of interpersonal synchrony: A meta-analysis. *Zeitschrift für Psychologie*, *224*(3), 168–189. https://doi.org/10.1027/2151-2604/a000252

Rogers, L. J., Wilson, K. G., Gohm, C. L., & Merwin, R. M. (2007). Revisiting written disclosure: The effects of warm versus cold experimenters. *Journal of Social and Clinical Psychology*, *26*(5), 556–574. https://doi.org/10.1521/jscp.2007.26.5.556

Rosenthal, R. (1963). On the social psychology of the psychological experiment: The experimenter's hypothesis as unintended determinant of experimental results. *American Scientist*, *51*(2), 268–283.

Rosenthal, R. (1973). *On the Social Psychology of the Self-Fulfilling Prophecy: Further Evidence for Pygmalion Effect and Their Mediating Mechanisms*. MMS Modular Publications.

Rosenthal, R. (1997). *Interpersonal Expectancy Effects: A Forty Year Perspective*. SAGE Publications.

Rosenthal, R. & Fode, K. (1963). Psychology of the scientist: V. Three experiments in experimenter bias. *Psychological Reports*, *12*, 491–511.

Rosenthal, R. & Jacobson, L. (1968). *Pygmalion in the Classroom: Teacher Expectation and Pupils' Intellectual Development*. Holt, Rinehart and Winston.

Rosenthal, R. & Rosnow, R. L. (2008). *Essentials of Behavioral Research: Methods and Data Analysis*, 3rd ed. McGraw-Hill.

Rosenthal, R. & Rosnow, R. L. (2009). *Artifacts in Behavioral Research*, 2nd ed. Oxford University Press.

Rosenzweig, S. (1933). The experimental situation as a psychological problem. *Psychological Review*, *40*(4), 337–354. doi:10.1037/h0074916

Saretsky, G. (1972). The OEO P.C. experiment and the John Henry effect. *The Phi Delta Kappan*, *53*(9), 579–581.

Shaywitz, S. E., Mody, M., & Shaywitz, B. A. (2006). Neural mechanisms in dyslexia. *Current Directions in Psychological Science*, *15*(6), 278–281. https://doi.org/10.1111/j.1467-8721.2006.00452.x

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22, 1359–1366. https://doi.org/10.1177/0956797611417632

Spring, B. & Alexander, B. L. (1989) Sugar and hyperactivity: another look. In R. Shepherd (ed.), *Handbook of the Psychophysiology of Human Eating* (pp. 231–249). Wiley.

Strickland, B. & Suben, A. (2012). Experimenter philosophy: The problem of experimenter bias in experimental philosophy. *Review of Philosophy and Psychology*, *3*(3), 457–467

Thorson, K. R., Mendes, W. B., & West, T. V. (2019). Controlling the uncontrolled: Are there incidental experimenter effects on physiologic responding? *Psychophysiology*, *57*, e13500. https://doi.org/10.1111/psyp.13500

Tuyttens, F. A. M., de Graaf, S., Heerkens, J. L. T., et al. (2014). Observer bias in animal behavior research: Can we believe what we score, if we score what we believe? *Animal Behaviour*, *90*, 273–280. http://dx.doi.org/10.1016/j.anbehav.2014.02.007

Vallier, H. & Timmerman, C. (2008). Clinical trials and the reorganization of medical research in post-Second World War Britain. *Medical History*, *52*(4), 493–510.

Vicente, K. J. & Brewer, W. F. (1993). Reconstructive remembering of the scientific literature. *Cognition*, *46*, 101–128.

Wason, P. C. (1968). Reasoning about a rule. *The Quarterly Journal of Experimental Psychology*, *20*(3), 273–281. https://doi.org/10.1080/14640746808400161

Weinstein, R. S. (2018). Pygmalion at 50: Harnessing its power and application in schooling. *Educational Research and Evaluation*, *24*(3–5), 346–365.

Winchester, C. L. & Salji, M. (2016). Writing a literature review. *Journal of Clinical Urology*, *9*(5), 309–312.

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, e1. https://doi.org/10.1017/S0140525X20001685

12 Debriefing and Post-Experimental Procedures

Travis D. Clark and Ginette Blackhart

Abstract

The steps social and behavioral scientists take after the end of a study are just as important as the steps taken before and during it. The goal of this chapter is to discuss the practical and ethical considerations that should be addressed before participants leave the physical or virtual study space. We review several post-experimental techniques, including the debriefing, manipulation checks, attention checks, mitigating participant crosstalk, and probing for participant suspicion regarding the purpose of the study. Within this review, we address issues with the implementation of each post-experimental technique as well as best practices for their use, with an emphasis placed on prevention of validity threats and the importance of accurate reporting of the steps taken after the experiment ends. Finally, we emphasize the importance of continuing to develop and empirically test post-experimental practices, with suggestions for future research.

Keywords*: *Debriefing, Manipulation Check, Crosstalk, Suspicion Probe, Attention Check, Instructional Manipulation Check

Introduction

Social and behavioral scientists have several decisions to make about the end of a study. What is the best way to properly debrief participants? How do we check that our data are accurate and uncompromised? How do we identify participants whose data were tainted and how should we treat those data? This chapter covers the practical and ethical considerations that should be addressed before participants leave the physical or virtual study space. These considerations include debriefing, probing for suspicion, safeguarding against crosstalk, and other issues. In this chapter, we critically examine the types of assumptions researchers commonly make about these post-experimental procedures and the available empirical evidence regarding those assumptions. Finally, we cover suggestions for best practices in administering these procedures.

Review of Procedures

The essential post-experimental procedure is a *debriefing* wherein the researcher explains the procedures that a participant underwent. On the surface, a debriefing should be administered at the conclusion of every study as it is an essential part of the code of ethics in many scientific organizations (see Chapter 2 in this volume). The American Psychological Association (APA) Code of Ethics, for example, has included debriefings since its inception in 1973 (see American Psychological Association, 2017, Section 8.07). According to these guidelines, a debriefing must be administered when deception is a component of the study. The American Sociological Association (2018) and the National Communication Association (2017) have similar provisions in their respective codes of ethics. The debriefing is both an ethical and practical concern; full disclosure is required for researchers to treat participants with dignity. Additionally, by encouraging participants not to discuss a study's hypotheses, a debriefing can protect the integrity of data from participant *crosstalk* (Edlund et al., 2009, 2014) (sometimes called inter-subject communication; Aronson, 1966) that occurs when participants give information to potential future participants. Crosstalk is typically detected using a suspicion probe, a procedure discussed at length in this chapter.

Most experiments should also include a *manipulation check* – an assessment of how well the experimental procedures manipulated the variable of interest. The term "manipulation check" is sometimes used interchangeably with the term *attention check* – a procedure designed to determine whether participants attended to the instructions or procedures used in the study. Attention checks can be failed because participants rush through a study for credit or pay but are also often failed by participants whose cognitive resources are depleted and their attention is flagging for that reason (Oppenheimer et al., 2009). In this chapter, we will use the term manipulation check for procedures that test the construct validity or effectiveness of an experimental manipulation of a variable whereas the term attention check will be used exclusively to refer to procedures detecting inattention.

Lack of attention can change participant responses in random ways, but a more insidious issue is suspicion. If participants are suspicious of study procedures, it can change their behavior (Blackhart et al., 2012). For this reason, a *suspicion probe* should be conducted at the conclusion of an experiment (Orne, 1962). A suspicion probe is typically introduced to detect participants' awareness from sources outside the experimental procedures but can also be useful for detecting knowledge or pressure created by procedures themselves. Such influence includes *demand characteristics*. Demand characteristics of the experimental situation are situational forces created by the study environment that push participants toward certain behaviors (Aronson et al., 1998). That is, participants subject to demand characteristics are more likely – consciously or unconsciously – to engage in behaviors that confirm the researchers' hypotheses because they are reacting to the situational pressure of being in an experiment (Orne, 1962). Contrary to the neat vocabulary presented here, these post-experiment procedures are often combined to such a degree that the boundary between each procedure is fuzzy in practice.

Why Focus on *Post* Experiment?

In an oft-cited research methods handbook chapter, Wilson et al. (2010, p. 123) point out four major goals of post-experimental procedures:

1. Ensure the well-being of participants.
2. Ensure that participants understood the experience and gained knowledge from it.
3. Use participants' perspective as a "consultant" in research to ensure the quality of study materials.
4. Probe for participant suspicion.

Indeed, the authors also state (p. 73) that "[i]t is impossible to overstate the importance of the post-experimental follow-up." Unfortunately, many researchers do not include the full details of these post-experimental procedures unless they happen to be related to the researchers' variables of interest (Ejelöv & Luke, 2020; Miketta & Friese, 2019). In fact, the same chapter contains another quote that has been empirically demonstrated to be false: "By [the end of the post-experimental inquiry], if deception has been used and any participants have any suspicions, they are almost certain to have revealed them" (Wilson et al., 2010, p. 76). There is a gap between the importance scientists place on post-experimental procedures and the care with which they are researched and reported.

There was a flurry of publications on the effects, limitations, ethical considerations, and recommended techniques for post-experiment procedures in the 1960s, when striking studies, such as the Milgram obedience to authority experiments (Milgram, 1963), gripped the public's consciousness (see Holmes, 1976, for a temporally proximal review). One culmination of this fear could be Tesch's (1977) article on the purposes of debriefing. Tesch's identification of the ethical, educational, and methodological purposes of debriefing still reflects the literature today (Tesch, 1977; see Sharpe & Faye, 2009, for a contemporary perspective). A lull then exists in the publication record, with relatively few empirical investigations of the subject until a renewed interest in the 2010s.

What explains this lull? One reason may be a change in the scope of what are considered post-experimental components. For example, when Ejelöv and Luke (2020) surveyed the literature to see what researchers were doing in their manipulation checks and how they were reporting them, they found a wide variety of procedures and an even wider range of definitions of what a manipulation check is and what counts as a successful manipulation check. Manipulation checks are such a universal feature in studies and such a logical choice for certain outcomes that they go unnoticed like a fish unaware it is swimming in water. Other critiques are presented below, but we believe there is a renewed interest in critiquing and improving the validity and replicability of scientific results. The last decade has seen the so-called Replication Crisis (see, for example, Edlund et al., 2022; Junk & Lyons, 2021; Open Science Collaboration, 2015; Zadvinskis & Melnyk, 2019), a renewed study of hypothesizing after the results are known (aka HARKing; Rubin, 2017), debate over null hypothesis statistical testing (Nuijten et al., 2016), and many more examples of the way we do science being put to the fire.

Like the scared motorist who refuses to have a mechanic look under the hood of their suspiciously loud car, scientists have discovered that empirically investigating the inner workings of social and behavioral science does lead to finding problems. Also like the unfortunate motorist, scientists know that discovering problems is the first step to finding solutions. Many researchers hope that increased scrutiny will ultimately improve and advance science (Edlund et al., 2022). In this chapter, you will find an introduction to several common post-experimental components, a review of the problems associated with each, and best practices for their implementation.

Manipulation Checks

Most scientists believe manipulation checks are necessary (Fayant et al., 2017). A manipulation check is primarily defined as a procedure that checks whether the manipulation in an experiment was successful (Hauser et al., 2018). Thus, we are talking about both internal and construct validity of the target manipulation. A manipulation check can occur after a study's conclusion, after a pilot design, or after an independent variable manipulation has been introduced (but before the dependent variable is collected). Early usage of the term "manipulation check" refers to an independent measure used to determine whether the experimental manipulation manipulated the intended variable (Wilson et al., 2010). As Ejelov and Luke (2020) observe, however, the usage of the term has expanded over time. Manipulation checks are a fuzzy concept, a fact not aided by their widely varying appearance within and across disciplines. According to their survey of the literature, it is likely that less than half of published manipulation checks refer to a measure of the independent variable (the original definition; Ejelov & Luke, 2020), with some authors using the term synonymously with "attention check" and other scientists using the term to refer to measures of mediating variables (Hauser et al., 2018). Inconsistencies in the way manipulation checks are reported was another common theme Ejelov and Luke (2020) discovered, with the level of detail often left up to guesswork.

Internal validity is the degree to which the observed effect in an experiment is due to the manipulation used. Features of an experiment's design all have the potential to systematically bias participant behavior. Elements of the design may make participants suspicious, nervous, or influence their cognitive resources, for example. One purpose of a manipulation check is to examine – preferably in participants' own words – what features of the experiment were salient. A manipulation check can provide evidence that the experimental manipulation was successful (or not) at influencing the construct it was intended to influence and that your manipulation is the sole (or at least primary) effect on that construct. Of course, this hinges on construct validity as well.

Construct validity is the degree to which a measurement instrument measures the construct it is intended to measure (Cronbach & Meehl, 1955). More pertinent to this chapter, construct validity can also refer to how much an experimental manipulation affects the construct it is intended to affect and not other, related, constructs. Threats to construct validity can be mundane (e.g., the experimenter may need to assess whether an anxiety-inducing experimental situation induced anxiety in participants) but can be potentially disastrous. If participants are suspicious, for example, their responses to measures may not reflect the construct under scrutiny but may instead result from reactivity to the experimental situation itself.

The use of manipulation checks to gain information about construct validity has critics (e.g., Sigall & Mills, 1998). When there are no working alternative explanations for the relationship between a manipulated variable and the dependent variable, a successful manipulation check may not provide any additional information. When there are plausible alternative explanations, a successful manipulation check shows that *something* happened in your experiment but does not rule out that the observed changes are due to changes in related constructs (not to the construct of interest). Ejelöv and Luke (2020) suggest assessing your construct of interest, related constructs, and unrelated constructs with a manipulation check depending on the particulars of the experiment. One or more of these may be appropriate. In this way, the manipulation check provides important divergent or convergent validity information for the true variable(s) of interest. There is no one-size-fits-all approach possible here; researchers must determine for each experiment which constructs it is acceptable for the target manipulation to causally influence and which constructs would be considered a failure of the intended manipulation.

When manipulation checks are reported, what do they typically look like? Not much systematic research has examined this question. Chester and Lasko (2021) coded two volumes of the *Journal of Personality and Social Psychology* and found that the most common manipulation check in this social psychology journal is a self-report scale of some kind, usually a one-item measurement. Multiple-item self-report scales are uncommon and behavioral tasks are rare; a surprising number of studies do not report the format at all (see Chester & Lasko, 2021, Figure 5). Hauser et al. (2018) found that approximately one-third of published studies report manipulation checks, but this varies by discipline. In the *Journal of Personality and Social Psychology,* some type of manipulation is employed in about two-thirds of published articles and approximately one-third of these studies employ a manipulation check (Chester & Lasko, 2021). For those studies without a manipulation check, the validity of the experimental manipulation is not directly checked or discussed in almost half of cases (of course, it is possible it was checked but this is not in the published record; Chester & Lasko, 2021).

Every researcher should critically think about what the manipulation check is for and what it can achieve. If the manipulation check fails, why is that? It could be because the independent variable was not measured correctly (Fayant et al., 2017), the check is insensitive (Blackhart et al., 2012), or because the manipulation did not work. One blind spot in reported psychological experiments is the decision making of researchers regarding what constitutes a failed manipulation check and why. Additional uses for the manipulation check are as an independent variable or to ascertain manipulation strength (for a full discussion, see Fayant et al., 2017).

Challenges

Order effects are a well-known problem in research, but the presentation of a manipulation check usually takes place at various points of a study (Ejelov & Luke, 2020). There is evidence that the position of the manipulation check in an experiment has an influence on certain dependent variables (Kühnen, 2010); it is probable that, if order effects were tested, more phenomena would be shown to be influenced by the presence of a manipulation check. Hauser et al. (2018) summarized several additional issues with manipulation checks as they are often employed. First, manipulation checks are an additional event that a participant experiences and can impact participants' behaviors. Feelings that participants experience can dampen or sharpen after administering a manipulation check that asks participants to reflect on those feelings (e.g., Keltner et al., 1993). Next, Hauser et al. (2018) discuss the measurement and analysis issues inherent in manipulation checks. Manipulation checks are estimates of the true strength and effectiveness of a manipulation, but many manipulation check procedures in practice offer a binary outcome, rather than a continuous measure (Chester & Lasko, 2021).

Best Practices

The most important recommendation for conducting a manipulation check is to report the details of the manipulation check as if it is an important part of the experimental procedures. Unfortunately, many studies fail to do so (Chester & Lasko, 2021). It is alarming that many researchers fail to include any validation information of their independent variable manipulations in published reports. A positive manipulation check can indicate that the manipulation in a study influenced a construct as intended, but a positive check could also be produced by artifacts of the experimental design or the measurement of related constructs (Chester & Lasko, 2021).

We introduced the potential problem above that the manipulation check procedure itself can cause changes in behavior at other points in an experiment. There are several ways to handle this issue – essentially a problem of order effects. First, several authors (e.g., Hauser et al., 2018) have suggested putting manipulation checks back where they were originally common – in pilot studies rather than (or in addition to) in a final experiment. We also recommend counterbalancing as a solution. Counterbalancing, in this case, would involve varying the presentation of the manipulation check to different points in the experiment and then examining for possible order effects.

Attention Checks

Attention checks are another useful way to check a manipulation's success. The problems that attention checks intend to detect include attention, distraction, careless responding, satisficing, or depleted cognitive resources. Participants may be particularly prone to inattention and distraction if they are performing tasks with the

https://doi.org/10.1017/9781009010054.013 Published online by Cambridge University Press

prominent purpose being a monetary one; this impacts the increasingly common use of services like Prolific Academic and Amazon Mechanical Turk. Some comparison studies indicate that problematic behaviors (e.g., inattention) are just as common in these online collection methods (Necka et al., 2016).

Challenges

Some studies show a failed attention check rate as high as 30% (e.g., Oppenheimer et al., 2009) for in-lab studies and 23% in online samples (e.g., Nichols & Edlund, 2020). Attention checks are inflexible in their implementation since researchers need to ensure participants are attending to the study's procedures at fixed points in many study designs. The simplest check of attention is a fact-based question with an obviously correct answer, such as "Which is bigger, the sun or the earth?"

Participants may "fail" this style of attention check because of their cultural background or their primary language spoken. For example, one student admitted to being concerned their data would be thrown out because they put the wrong answer to the attention check question "Who is the current president?" We recommend carefully designing attention check questions so that they require no specific cultural background and can be answered easily by *any* speaker of the study's primary language. Using simple, easy-to-read language will also benefit participants low in attentional resources.

Participants genuinely get these types of simple questions wrong. Incorrect answers on questions such as this are perhaps due to satisficing (Oppenheimer et al., 2009). Satisficing occurs when participants find an answer on a task that is "good enough" rather than expending full cognitive effort to find the correct answer. On our sun/earth example, we imagine at least some additional participants answered indiscriminately – after all, participants have limited time to complete the survey and, with only two options, one of them must be correct.

It is worth noting that there is a popular notion that Mechanical Turk samples are less attentive than other samples (discussed in Chandler et al., 2014). Some online crowd-based samples are comparable in inattention with physical samples (e.g., Necka et al., 2016) and comparable on other measures of data quality (Kees et al., 2017). Evidence gathered with an instructional manipulation check (see below) suggests Mechanical Turkers are more attentive than undergraduate participant pools (Hauser & Schwarz, 2016). Until the pattern of evidence is clear and the conditions that promote data quality in online samples are well defined, further comparisons are necessary.

Best Practices

One alternative to such attention check items is to provide the answer to participants in the question. For example, a researcher could ask the following question: "Which state within the United States of America was the last state to be admitted to the union? (The answer is Hawaii)" with the response options of Alaska, Hawaii, Puerto Rico, Arizona, and New Mexico. This will eliminate concerns about cultural background, language, and satisficing influencing how participants answer these types of attention check items. Another, and perhaps better, solution is to format attention check questions in what Oppenheimer et al. (2009) refer to as an *instructional manipulation check* or what Meade and Craig (2012) refer to as an *instructed response item*.

The instructional manipulation check or instructed response item is a question embedded in the study materials that takes the same format as the other study materials but asks participants to ignore the instructions and enter a specific answer. In their example, Oppenheimer et al. (2009) provide a "check all that apply" type question that has embedded instructions to skip the question entirely and instead click on the question title to proceed to the next page. These instructions subvert all expectations of taking a survey online, so participants who do click the question title undoubtedly have their attention focused on reading the instructions in the survey. Another example would be to use an item with explicit instructions on which answer participants should select, such as, "Paying attention and reading the instructions carefully is critical. If you are paying attention, select '4' below." An instructional manipulation check also filters out participants who are not paying attention to the procedures for other reasons (e.g., to quickly finish a study for credit). The instructional manipulation check increases statistical power by discriminating participants who show poor survey performance (e.g., not recognizing reverse coded items, moving incredibly quickly through survey pages) from participants who appear to be genuinely answering questions (Oppenheimer et al., 2009).

Another consideration is that some participants may purposefully answer some of these questions incorrectly. For instance, Meade and Craig (2012) suggested that some participants might have purposefully answered some of their bogus questions (e.g., "All of my friends say I would make a great bicycle") incorrectly because they thought it was humorous to do so. For this reason, Meade and Craig recommend using the instructed response items over bogus items as attention check questions.

Kees et al. (2017, p. 155) offer several suggestions for best practices in online data collection that we believe apply to the use of attention checks more generally; for instance, implementing multiple attention checks throughout a study and "speed trap" items that require participants to spend a minimum time on a page that requires reading. Also, before collecting data, a research team can predetermine the acceptable level of threats to data integrity (such as evidence of careless responding or failed manipulation checks). In addition, Kees and colleagues recommend using a "soft launch" of your survey to collect a small amount of data, and then checking the survey and data for unexpected errors. This suggestion complements our philosophy of using participants as experts in study participation and using their feedback to improve study procedures. A "soft launch" of a study paired with a funnel debriefing procedure (described below) could solve many data collection issues.

Debriefing

The debriefing of an experiment is an ethical necessity for several reasons. A thorough debriefing is an educational experience for participants, helping reinforce their trust in the scientific process and increasing their positive feelings toward social and behavioral science. The explicit aim of many required undergraduate student research participation requirements is for the students to be exposed to research and to learn more about the research process (Zannella et al., 2020). Reviews of students' reactions after participating in research find that they do indeed report learning from the process (Zannella et al., 2020) and we believe this step is important for any participant (not just undergraduate students!). Many studies use deception or deceive participants by omitting information. In these cases, it is ethically necessary to both dehoax and desensitize participants.

Dehoaxing is the process wherein the experimenter reveals the true purpose of the study and all deception perpetrated (see Holmes, 1976). Desensitization is the component of the debriefing where the experimenter attempts to minimize any negative feelings participants may have experienced by participating in the study (see Holmes, 1976). Participants may experience negative feelings as an intentional part of the research design (e.g., failure feedback, ostracism inductions) or due to being deceived (e.g., shame, embarrassment). Participants may themselves have performed ethically questionable tasks (e.g., administering a shock or noise blast to another participant). According to the APA Code of Conduct, participants must learn the truth behind any deception as soon as possible (see American Psychological Association, 2017, Section 8.07), and it is recommended for this to take place during the debriefing process, when possible (see American Psychological Association, 2017, Section 8.08).

Debriefing and post-experiment questions also offer several practical advantages. In a classic investigation of demand characteristics, Orne (1962) was only able to accurately describe the phenomenon by following up with participants after the study procedures to investigate their frame of mind and thoughts while participating in the experiment. Demand characteristics are an important subject to consider for practical and ethical reasons (see Chapter 11 in this volume). The debriefing process can help identify when participants were pushed toward certain behaviors – a confound in a study design. Debriefing can also be used to assure participants that their behaviors in a study were influenced by the experimental situation and may not reflect their behaviors in real life. Additionally, participants' questions and concerns can be used to identify improvements to a study design. An unanticipated environmental cue (e.g., an uncomfortable couch, a difficult-to-read web page, or noise from an adjacent room) can easily become a confounding variable when testing subtle psychological manipulations, but participants can inform the researchers of unexpected issues.

Debriefing is also an important time to stress to participants that details of the study should not be shared with other individuals who may potentially participate in the study. In a typical physical setting, the college campus, details of psychological research can spread like wildfire through participant crosstalk (see below).

Challenges

It is often assumed that a thorough debriefing will protect participants from any lasting negative effects of an experiment, an assertion that is often embedded in our consent forms and ethics board applications. Real-life situations are different than the imagined experimental situation in this regard. Miketta and Friese (2019) summarize several empirical studies showing that negative feedback persists even after it is retracted; you cannot unring a bell. Women informed of a false-positive breast cancer diagnosis, for example, have increased health vigilance months later (Lerman et al., 1991).

In a series of experiments, Miketta and Friese (2019) investigated after-effects of the type of brief negative feedback often used in ego threat research. After failure feedback (being told their performance on an intelligence task was low) or social rejection feedback (being told that other participants found them unlikeable), participants' mood was assessed pre- and post-debriefing. The authors used a *revised outcome debriefing* (McFarland et al., 2007) and extended this debriefing with several additions; however, negative mood was not mitigated. Miketta and Friese resorted to an intensive procedure that mitigated some, but not all, of the negative mood effect Miketta and Friese 2019, pp. 304–305):

> [Participants] received a 10–15 min-long, extensive debriefing that was designed to address both potential cognitive and affective perseverance of the false feedback. This extensive debriefing was conveyed in a sensitive, caring, and emotionally warm manner by a carefully trained experimenter. A final debriefing version additionally informed participants about the perseverance effect, its potential significance for well-being after an ego threatening experience and instructed participants to take countermeasures if they noticed such perseverance on their well-being (extensive process debriefing).

Despite this intensive procedure, negative mood effects were not eliminated (although the decrease was statistically significant) two weeks after the study's conclusion.

Participant perceptions of the debriefing process itself are cause for concern, even if there was no deception or negative-mood-inducing manipulation. When Brody et al. (2000) coded the answers to several open-ended questions about one institution's debriefing practices, several problems stood out. The most alarming pattern is the low number of participants who left the study believing the debriefing was performed well (40.6%) and the number who believed the debriefing was unclear (28.8%). The most worrisome answer among participants is that experimenters often give them no debriefing or minimal debriefing (Brody et al., 2000). Empirical investigations often find that many authors do not report debriefing participants, but it is often assumed that the scientists are debriefing but omitting this process from their method sections (see Brody et al., 2000; Sharpe & Faye, 2009; Zannella et al., 2020). However, when pressed for additional information, some scientists do admit providing *no debriefing at all* (Sharpe & Faye, 2009), with the rationale that no deception was involved in the study.

Zannella et al. (2020) assessed undergraduate students' educational experiences upon participating in research and gave students an opportunity to provide feedback on how to improve their educational outcomes. They found that, on average, students have positive experiences participating in research. However, what leads some participants to have bad experiences? The most relevant finding to this chapter is that students report that the debriefing of experiments is unclear or too short to be informative.

Best Practices

Best practices for debriefing are difficult to ascertain. This is due to a combination of methodological diversity and lack of reporting in published studies (Miketta & Friese, 2019; Sharpe & Faye, 2009). Reflecting on this, our first suggestion is for authors to carefully report debriefing procedures and journal editors to request this information. Best practices should include ethical, methodological, and educational aims (Sharp & Faye, 2009; Tesch, 1977).

For ethical reasons and also to ensure methodological integrity, we support the suggestion of Miketta and Friese (2019) that debriefing procedures and empirical checks of debriefing efficacy should be standard reporting after participants undergo an adverse event including deception. Miketta and Friese (2019, p. 306) ask three important questions of future research:

1. Which aftereffects of participating in psychological research do we, as a discipline, regard as acceptable versus problematic? . . .
2. How long do affective aftereffects caused by an ego threatening experience last, and what can be considered an ethically acceptable time frame? . . .
3. Are there real-life consequences of negative aftereffects caused by ego threatening experiences? . . .

Although their focus was on one phenomenon, an ego-threatening experience, these guidelines work as a thoughtful starting point for other areas of research. Researchers often have a cavalier approach to reporting ethical decision making in employing deception (Ortmann & Hertwig, 2002); we believe that the administration of a debriefing is similarly not given enough scrutiny.

To protect the data integrity of our colleagues who frequently use deception, we believe best practices for debriefing after deception should apply to all studies. Rates of deception in social and behavioral science journals can be as high as one out of two published studies in certain volumes (Adair et al., 1985). Wilson et al. (2010) suggest explaining to participants the damage that could be done to the scientific integrity of a study should information be spread to other participants, and giving participants an easy to remember, vague cover story about what an experiment is about. In our own research, we provided participants with a description of the experiment, emphasizing its goal to find group differences in decision making. This explanation of our study was accurate but vague; participants were informed of the details of the study in our debriefing procedures but asked to share the publicly available and vague study description if pressured for information from potential participants.

To combat the perseverance of negative effects following false feedback, we recommend a combination of a process debriefing approach with additional bogus feedback information (discussed in McFarland et al., 2007). Process debriefing involves outlining to participants the psychological processes that may impact them after receiving false feedback – namely, the perseverance of negative effects. We recommend explaining to participants the false nature of the feedback and the bogus nature of procedures used (McFarland et al., 2007; Miketta & Friese, 2019). It is not enough to know the feedback they received is false – it is also important that participants know that the tasks used do not truly measure the trait, characteristic, or ability about which they received feedback.

In an ideal situation, participants in a study can be essential to critiquing and improving the study (Wilson et al., 2010). We have a standard set of debriefing procedures that can be modified to fit the needs of diverse experiments. These procedures start with the prompt below, or similar language, to persuade participants to help the researcher(s) by critiquing the procedures:

> *We would like your feedback about the design of the study to be sure that our experimental design is sound. We want to know whether anything odd or irregular happened as you participated in the study today. These things sometimes happen and, if we know about them, we can correct for them and make sure that our findings are valid and reliable. It is therefore extremely important for the scientific validity of the study that you tell us whether anything like this happened today. Please be as honest as possible in your answers; no feedback we receive, including negative feedback, will result in a loss of [research credit/payment], nor will it affect how we use your data. In fact, negative feedback is an important way for us to improve upon our design for future studies. Be as detailed as you feel is necessary to fully answer each question. You may spend as much time on these questions as you want, but we ask that you spend a minimum of 5 minutes answering these questions.*

The wording of this prompt is the culmination of several objectives, each discussed throughout this chapter. First, participants are unlikely to disclose information that makes them seem like a "bad participant," so they are encouraged to provide potentially damaging information. The scientific integrity of the study is also emphasized to encourage honest feedback. Finally, as many participants are fatigued or inattentive by the end of a study, we recommend giving participants a time minimum to answer follow-up questions. We use a funnel debriefing procedure (see Blackhart et al., 2012) at the end of all procedures in our laboratories. Funnel debriefing starts with general questions and continues with specific questions. As more specific questions about the study may elicit more participant awareness or suspicion, it is important that participants should not be allowed to return to previous questions to edit their responses.

The more "traditional" debriefing and suspicion probe procedures take place in a laboratory where an experimenter or assistant can be a welcoming presence to establish rapport and to elicit feedback and questions. It is recommended that experimenters gently introduce participants to any deception and manipulation because people, understandably, do not like being manipulated (Wilson et al., 2010). Social and behavioral scientists should be keenly aware to avoid leading questions that might trigger participants to expect deception or plant ideas about the experiment before gaining feedback from them. One advantage of a funnel debriefing procedure is that it can be administered electronically without an experimenter's guidance.

Administering self-guided procedures at a computer or over the internet is a wonderful way to reduce biases (e.g., expectancy effects) but may introduce other sources of difficulty (e.g., reducing the personal connection between an experimenter and the participant). Much of the advice about administering debriefings, and especially suspicion probes, focuses on establishing a rapport with participants to encourage them to be open with feedback (Wilson et al., 2010), but this is obviously not possible when no live human presence is in the procedure. As we will discuss below, some investigations find more candid feedback from participants when the experimenter is not present (Blackhart et al., 2012); therefore, the available empirical literature does not support the advantage of experimenter presence.

We have discussed best practices from an ethical and methodological perspective, but the value of debriefings as an educational tool should not be overlooked. We have reviewed research finding that some participants view debriefings as useless and leave studies feeling less confident about science (e.g., Brody et al., 2000), but what can the literature tell us about participants who did find their participation meaningful and educational? Many organizational ethics codes cited in this chapter have educational value as an explicit goal of the debriefing process. Educational value is an important benefit of participating in research (Zannella et al., 2020) and participants themselves desire their participation to be informative and educational (Brody et al., 2000). Perhaps more specific to an undergraduate participation pool, many students enjoy participating in research that connects to topics they have learned about (Zannella et al., 2020).

In addition to educational value, we must address the practical value of making research more interesting. Participants undoubtedly share information about their experiences in studies, especially in online environments (e.g., Mechanical Turk; Edlund et al., 2017). Increasing one's reputation as a provider of research tasks is one important way of increasing data quality by recruiting participants with genuine interest in advancing science (Kees et al., 2017). Scientists should also safeguard the public's perceptions of science. By providing informative debriefing, we do our small (and required!) part to increase trust in science. Zannella et al. (2020) found that many participants wish to know the *results* of studies that they participate in and suggest adding a way for participants to provide an email address to learn the details of a study after data collection has stopped (when appropriate).

Participant Crosstalk

Participant crosstalk occurs when participants share information about a study with potential future participants. Contamination of the participant pool is more widespread than researchers realize. Early studies involved providing participants in a study with unique information – such as a "new" drug testing method (Diener et al., 1972) – then later probing a subject pool for the information. A more recent investigation provided participants with unique information (the number of gumballs in a glass jar), got verbal agreement to keep the information secret, then tested future participants for the information (Edlund et al., 2014).

Challenges

Rates of crosstalk in these studies vary across time and institution from minimal (0.5% or less) to problematic (above 3%), depending on factors such as institutional policies (Edlund et al., 2014). Online crowdsourced data collection methods, such as Amazon's Mechanical Turk, pose even greater challenges to participant naiveté. "Workers" have a network of websites to share information about worthwhile human intelligence tasks on social media (e.g., Reddit, Facebook) and on sites designed specifically for workers (e.g., mturkforum.com; see Chandler et al., 2014; Edlund et al., 2017). These websites provide online communities where workers can share information and recommendations beneficial to the worker community and are often used to share sensitive information about tasks (Edlund et al., 2017). Crosstalk in one sample involved 33% of studies discussed online (Edlund et al., 2017).

As previously mentioned, participant crosstalk is typically assessed with a suspicion probe. Unfortunately, several interventions have been found to be ineffective. The introduction of an incentive – in the form of extra research participation credit or cash – has mixed efficacy (Blackhart et al., 2012) and we caution against offering such rewards. To neutralize the effects of implicit norms participants may have about admitting information, Clark (2013) manipulated the presence of a norm prompt, indicating to participants that previous participants had been very open with feedback (particularly negative feedback) about study procedures. This and other post-experimental inquiry variations all failed to have a strong impact on post-experiment honesty rates. Although the presence of the norm prompt appeared to increase post-experimental inquiry accuracy, the low base rates of participants' admitted suspicions make it difficult to conclude whether this finding was a true positive result.

The authors of this chapter have experimentally manipulated several variables to test their effects on admission rates (Blackhart et al., 2012; Clark, 2013; Edlund et al., 2014) with some limited success. Comparing an in-person interview with an anonymous computer questionnaire, participants admitted to more suspicion and awareness of the study goals in an anonymous setting (Blackhart et al., 2012); pre-computer research found higher admission rates using pencil-and-paper survey than in in-person interviews (Newberry, 1973). Other variations on response type indicate that allowing more flexibility in reporting suspicion and awareness leads to greater positive reporting; in other words, let participants report whether they are *a little suspicious* or *moderately suspicious* rather than forcing them into a binary choice (Newberry, 1973). An interview and open-ended inquiry achieves this goal, for example, where a checkbox indicating suspicion would not.

Best Practices

Crosstalk cannot be prevented entirely. Early efforts to reduce crosstalk involved explaining to participants how crosstalk could affect scientific integrity (Aronson, 1966; Golding & Lichtenstein, 1970). In addition, a signed statement of confidentiality appears to reduce crosstalk (Walsh & Stillman, 1974). Furthermore, Edlund and colleagues introduced the following simple prompt: "I would like to ask that you not tell anyone about this experiment to help keep guesses normal. Is that okay with you?" (see Edlund et al., 2014; originally reported by Edlund et al., 2009).

A similar request not to share information was used by Edlund and colleagues in a Mechanical Turk sample that successfully reduced crosstalk rates (Edlund et al., 2017). Our laboratories use a more detailed prompt for all experiments. It is important to standardize a prompt for studies so that participants in a particular pool (e.g., university participant pools; crowdsourcing participant pools) are not aware that deception is used simply by the presence of the scientific integrity prompt. Before completing post-experimental inquiries, participants are given the prompt found in the best practices detailed in the Debriefing section above.

Before beginning data collection for an experiment, we recommend discussing participant crosstalk with other researchers who may share your participant pool (e.g., members of your department or institution if at a university). There may be a link between stringent institutional research participation requirements and increased crosstalk, although this relationship needs to be explored further (Edlund et al., 2014). We also recommend variation and testing of post-experiment procedures at the individual study and participant pool levels. Debriefing and post-experimental procedures are used by many researchers but are rarely systematically studied for efficacy. More research is needed in this area as the current research indicates procedures may vary in efficacy by institution and for different procedures (Edlund et al., 2014).

Suspicion Probes

Participants' expectations about an experiment's hypotheses can guide their behavior in unnatural ways (Orne, 1962). This issue is particularly troublesome when deception is used, as some participants may see through the cover story of a study or false feedback given to them by an experimenter. Deception thus vastly increases the necessity for a suspicion probe, but we argue that a suspicion probe is relevant in many procedures even when deception is not used. A suspicion probe is any procedure designed to assess participant suspicion of the true nature of a study's procedures and awareness of the hypotheses and aims of the experiment that the experimenter has kept obscured. In many cases, it is also helpful to assess participant awareness of study goals that the experimenter has provided to the participant, such as those provided in the informed consent procedures. It is often assumed that participants will understand the instructions given to them, which is both a practical and ethical necessity, but a thorough suspicion probe may reveal otherwise. As discussed above, studies indicate that it is common for participants in a participant pool or on an online crowdsourcing site to engage in crosstalk (see Edlund et al., 2009, 2017; Lichtenstein, 1970).

Challenges

Golding and Lichtenstein (1970) present four factors necessary for the legitimate use of deception in research: (1) suspiciousness of the study protocol will not affect its response outcomes, (2) participants arrive with little or no knowledge of the study, (3) the study does not indicate to participants that they are being deceived, and (4) knowledge of the study gained before or during the protocol can be assessed by the experimenter. Most researchers using deception explicitly or implicitly make these four assumptions but, as we will review, each of them must be carefully considered and not assumed to be true.

Suspiciousness of the study protocol affects participant responses in unexpected ways. Participants who participate in a study in which they are deceived are more suspicious in the future and their behavior may change as a result (Cook & Perrin, 1971; Hertwig & Ortmann, 2008). Participants in a negative mood state are more likely to be suspicious of a study; combine this with the number of deceptive procedures that elicit negative experiences (e.g., social rejection) and it becomes a serious methodological concern (Forgas & East, 2008). Another specific behavior that decreases when participants are suspicious is conformity (Hertwig & Ortmann, 2008). Some participants may be suspicious of the study design, while others may directly know the hypotheses.

Ideally, suspicion probes will discover these nuances. Unfortunately, the efficacy of suspicion probes may be lacking. Several researchers have given information about a study to participants via an experimental confederate and found that none of the participants revealed having information in a post-experimental inquiry (McMillen & Austin, 1971; Nichols & Maner, 2008). When the participants in one study pooled together their information and uncovered the deception of the procedures, none admitted this on the post-experimental inquiry (Taylor & Sheppard, 1996). Other researchers have more hopeful results, such as one participant (out of 81; Sagarin et al., 1998) revealing that they knew about the task from a confederate or one participant (out of 16; Levy, 1967) admitting to receiving information before the study began. There is clear evidence that the assumption that non-naïve participants will reveal their suspicions or knowledge is suspect.

Blackhart et al. (2012) introduced knowledge of an experiment as an independent variable with a confederate randomly assigned to supply participants with information about the study (i.e., half of the participants were informed of the supposed true purpose of the study, but the other half did not receive such information). If Wilson et al.'s (2010) assertion is true, that participants at the end of a debriefing process will reveal their knowledge of a study, we would expect approximately half of participants to admit to having prior information. In the study by Blackhart et al. (2012) and a follow-up (Clark, 2013), less than one in five "spoiled" participants revealed their information.

Blackhart et al. (2012) surveyed an undergraduate participant pool to discover possible reasons for participants' reluctance to reveal information. The most common answers among this participant pool were concerns about ruining the study (44%), concerns about not receiving research credit or payment (39%), concerns about getting someone in trouble (31% expressed concern about getting themselves into trouble and 35% expressed concern about getting someone else into trouble), and concerns about feeling foolish (26%).

Best Practices

Many researchers do not include suspicion probes or only include vague descriptions of their suspicion probe (Chester & Lasko, 2021). Lack of reporting or vague reporting has led to an accumulation of best practices from individual researchers without good discipline-wide recommendations in the social and behavioral sciences. We will describe some of our procedures below but recommend a thorough review of *prevention* practices in our Crosstalk section before considering different *detection* practices from the current section.

Suspicion probe data are questionable and need further empirical investigation but are undoubtedly still the best way to divide naïve participants from non-naïve participants. If participants do indicate "genuine awareness" of a study's procedures, the most straightforward step is to remove their data from data analyses (suggested by Bargh & Chartrand, 2000). This course is common but may introduce a systemic bias into the data (Shimp et al., 1991). For instance, participants who are aware of the procedures could vary in particular ways from participants who are not aware – they may be more intelligent, have a higher need for cognition, or have greater attentional resources while completing the procedures (Shimp et al., 1991). Removing participants without investigating potential differences could jeopardize the construct, internal, and external validity of the study. Shimp et al. (1991) recommend performing a sensitivity analysis in which statistical analyses are run and reported with and without the aware participants. If the main analyses change depending on participant awareness, we recommend, at a minimum, reporting both sets of results so the reader can draw their own conclusions about the inconsistencies. An additional step would be to perform a conceptual replication of ones' own study with intentional manipulation of participant awareness levels to provide further evidence that the difference between aware and naïve participants is truly due to awareness rather than to other confounding variables.

What about suspicious participants who fall below that threshold of "genuine awareness"? Current practices are varied. Some researchers choose to discard data from participants expressing suspicion above a certain threshold whereas others statistically control for suspicion to determine the degree to which it influences relevant outcomes. We will reiterate our suggestion of assessing suspicion in a way that allows some variability. While participants are reluctant to agree to knowledge of a study due to concerns of compromising the data (e.g., appearing to be a bad participants), they are more likely to express some gradient of suspicion or awareness of the study protocols. Best practices in the social and behavioral sciences include the selection of inclusion or exclusion criteria before an experiment begins. Similarly, researchers should determine what level of suspicion should disqualify a participant from data analysis before any data are collected (Wilson et al., 2010). Although we agree that a priori decision making should be performed in most cases, we have discovered truly surprising participant suspicions upon data collection that would never have fit into our preconceived notions of what participants may guess about a particular study. One potential solution is to modify one's exclusion criteria after data have been collected or to perform data analyses with and without suspicious participants to determine, for a particular set of variables, whether suspicion impacts behavior. Congruent with suggestions made above, pilot testing is another excellent way to determine, ahead of time, what types of information participants may (correctly or incorrectly) glean from your procedures.

Conclusion

We must end this chapter by reiterating the importance of empirically testing the post-experimental practices presented here. Researchers should empirically test various post-experimental practices, add these practices to the scientific record, and add an ethical decision-making rationale to manuscripts (see Chapter 2 in this volume). Too much of the current literature in various social and behavioral science disciplines offers scant information about their debriefing protocols. This will only change with a culture shift toward introspective examination of often unexamined experimental practices.

Further Reading

For a detailed example of a funnel debriefing procedure and the empirical test of various post-experimental practices including suspicion probing, we recommend the following article:

Blackhart, G. C., Brown, K. E., Clark, T., Pierce, D. L., & Shell, K. (2012). Assessing the adequacy of postexperimental inquiries in deception re-search and the factors that promote participant honesty. *Behavior Research Methods*, *44*, 24–40. https://doi.org/10.3758/s13428-011-0132-6

For further discussion of the history and progression of manipulation checks as well as specific recommendations for their use, we recommend Table 4 in the following article:

Ejelöv, E. & Luke, T. (2020). "Rarely safe to assume": Evaluating the use and interpretation of manipulation checks in experimental social psychology. *Journal of Experimental Social Psychology*, *87*, 103937. https://doi.org/10.1016/j.jesp.2019.103937

We are proponents of manipulation checks (with the proper precautions), but criticisms of manipulation checks should be seriously considered. For further reading on critiques of manipulation check practices we recommend the following article:

Hauser, D., Ellsworth, P., & Gonzalez, R. (2018). Are manipulation checks necessary? *Frontiers in Psychology*, *9*, 998. https://doi.org/10.3389/fpsyg.2018.00998

References

Adair, J., Dushenko, T., & Lindsay, R. (1985). Ethical regulations and their impact on research practice. *The American Psychologist*, *40*, 59–72. https://doi.org/10.1037//0003-066X.40.1.59

American Psychological Association (2017). Ethical principles of psychologists and code of conduct (2002, amended effective June 1, 2010, and January 1, 2017). Available at: www.apa.org/ethics/code.

American Sociological Association (2018). Code of ethics. Available at:www.asanet.org/sites/default/files/asa_code_of_ethics-june2018a.pdf.

Aronson, E. (1966). Avoidance of inter-subject communication. *Psychological Reports*, *19*, 238. https://doi.org/10.2466/pr0.1966.19.1.238

Aronson, E., Wilson, T. D., & Brewer, M. B. (1998). Experimentation in social psychology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (eds.), *The Handbook of Social Psychology* (pp. 99–142). McGraw-Hill.

Bargh, B. A. & Chartrand, T. L. (2000). The mind in the middle: A practical guide to priming and automaticity research. In H. T. Reis & C. M. Judd (eds.), *Handbook of Research Methods in Social and Personality Psychology* (pp. 253–285). Cambridge University Press.

Blackhart, G. C., Brown, K. E., Clark, T., Pierce, D. L., & Shell, K. (2012). Assessing the adequacy of postexperimental inquiries in deception research and the factors that promote participant honesty. *Behavior Research Methods*, *44*, 24–40. https://doi.org/10.3758/s13428-011-0132-6

Brody, J. L., Gluck, J., & Aragon, A. S. (2000). Participants' understanding of the process of psychological research: Debriefing. *Ethics and Behavior*, *10*, 13–25, https://doi.org/10.1207/S15327019EB1001_2

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*, 112–130. http://dx.doi.org/10.3758/s13428-013-0365-7

Chester, D. S. & Lasko, E. N. (2021). Construct validation of experimental manipulations in social psychology: Current practices and recommendations for the future. *Perspectives on Psychological Science*, *16*, 377–395. https://doi.org/10.1177/1745691620950684

Clark, T. D. (2013). Using social influence to enhance post-experimental inquiry success (unpublished Master's thesis). University of North Dakota, Grand Forks, ND.

Cook, T. D. & Perrin, B. F. (1971). The effects of suspiciousness of deception and the perceived legitimacy of deception on task performance in an attitude change experiment. *Journal of Personality*, *39*, 204–224. https://doi.org/10.1111/j.1467-6494.1971.tb00037.x

Cronbach, L. & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302. https://doi.org/10.1037/h0040957

Diener, E., Matthews, R., & Smith, R. E. (1972). Leakage of experimental information to potential future subjects by debriefed participants. *Journal of Experimental Research in Personality*, *6*, 264–267.

Edlund, J. E., Sagarin, B. J., Skowronski, J. J., Johnson, S., & Kutter, J. (2009). Whatever happens in the laboratory stays in the laboratory: The prevalence and prevention of participant crosstalk. *Personality and Social Psychology Bulletin*, *35*, 635–642. https://doi.org/10.1177/0146167208331255

Edlund, J. E., Nichols, A. L., Okdie, B. M., (2014). The prevalence and prevention of crosstalk: A multi-institutional study. *The Journal of Social Psychology*, *154*, 181–185. https://doi.org/10.1080/00224545.2013.872596

Edlund, J. E., Lange, K. M., Sevene, A. M., et al. (2017). Participant crosstalk: Issues when using the Mechanical Turk. *Tutorials in Quantitative Methods for Psychology*, *13*, 174–182. http://doi.org/10.20982/tqmp.13.3.p174

Edlund, J. E., Cuccolo, K., Irgens, M. S., Wagge, J. R., & Zlokovich, M. S. (2022). Saving science through replication studies. *Perspectives on Psychological Science*, *17*(1), 216–225. https://doi.org/10.1177/1745691620984385

Ejelöv, E. & Luke, T. (2020). "Rarely safe to assume": Evaluating the use and interpretation of manipulation checks in experimental social psychology. *Journal of Experimental Social Psychology*, *87*, 103937. https://doi.org/10.1016/j.jesp.2019.103937

Fayant, M.-P., Sigall, H., Lemonnier, A., Retsin, E., & Alexopoulos, T. (2017). On the limitations of manipulation checks: An obstacle toward cumulative science.

International Review of Social Psychology, *30*, 125–130. https://doi.org/10.5334/irsp.102

Forgas, J. P. & East, R. (2008). On being happy and gullible: Mood effects on skepticism and the detection of deception. *Journal of Experimental Social Psychology*, *44*, 1362–1367. https://doi.org/10.1016/j.jesp.2008.04.010

Golding, S. L. & Lichtenstein, E. (1970). Confession of awareness and prior knowledge of deception as a function of interview set and approval motivation. *Journal of Personality and Social Psychology*, *14*, 213–223. https://doi.org/10.1037/h0028853

Hauser, D. J. & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*, 400–407. https://doi.org/10.3758/s13428-015-0578-z

Hauser, D., Ellsworth, P., & Gonzalez, R. (2018). Are manipulation checks necessary? *Frontiers in Psychology*, *9*, 998. https://doi.org/10.3389/fpsyg.2018.00998

Hertwig, R. & Ortmann, A. (2008). Deception in experiments: Revisiting the arguments in its defense. *Ethics and Behavior*, *18*, 59–92. https://doi.org/10.1080/10508420701712990

Holmes, D. S. (1976). Debriefing after psychological experiments: I. Effectiveness of post-deception dehoaxing. *American Psychologist*, *31*, 858–867. https://doi.org/10.1037/0003-066X.31.12.858

Junk, T. R. & Lyons, L. (2021). Reproducibility and replication of experimental particle physics results. *PsyArXiv*. https://arxiv.org/abs/2009.06864.

Kees, J., Berry, C., Burton, S., & Sheehan, K. (2017). An analysis of data quality: Professional panels, student subject pools, and Amazon's Mechanical Turk. *Journal of Advertising*, 46, 141–155. https://doi.org/10.1080/00913367.2016.1269304

Keltner, D., Locke, K. D., & Audrain, P. C. (1993). The influence of attributions on the relevance of negative feelings to personal satisfaction. *Personality and Social Psychology Bulletin*, *19*, 21–29. https://doi.org/10.1177/0146167293191003

Kühnen, U. (2010). Manipulation checks as manipulation: Another look at the ease-of-retrieval heuristic. *Personality and Social Psychology Bulletin*, *36*, 47–58. https://doi.org/10.1177/0146167209346746

Lerman, C., Trock, B., Rimer, B. K., et al. (1991). Psychological side effects of breast cancer screening. *Health Psychology*, *10*, 259–267. https://doi.org/10.1037/0278-6133.10.4.259

Levy, L. (1967). Awareness, learning, and the beneficent subject as expert witness. *Journal of Personality and Social Psychology*, *6*, 363–370.

Lichtenstein, E. (1970). "Please don't talk to anyone about this experiment": Disclosure of deception by debriefed subjects. *Psychological Reports*, *26*, 485–486.

McFarland, C., Cheam, A., & Buehler, R. (2007). The perseverance effect in the debriefing paradigm: Replication and extension. *Journal of Experimental Social Psychology*, 43, 233–240. https://doi.org/10.1016/j .jesp.2006.01.010

McMillen, D. & Austin, J. (1971). Effect of positive feedback on compliance following transgression. *Psychonomic Science*, *24*, 59–61. https://doi.org/10.3758/BF03337892

Meade, A. W. & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Miketta, S. & Friese, M. (2019). Debriefed but still troubled? About the (in)effectiveness of postexperimental debriefings after ego threat. *Journal of Personality and Social Psychology*, *117*, 282–309. https://doi.org/10.1037/pspa0000155

Milgram, S. (1963). Behavioral study of obedience. *The Journal of Abnormal and Social Psychology*, *67*, 371–378. https://doi.org/10.1037/h0040525

National Communication Association (2017). A code of professional ethics for the communication scholar/teacher. Available at: www.natcom.org/sites/default/files/pages/1999_Public_Statements_A_Code_of_Professional_Ethics_for_%20the_Communication_Scholar_Teacher_November.pdf.

Necka, E., Cacioppo, S., Norman, G., & Cacioppo, J. (2016). Measuring the prevalence of problematic respondent behaviors among MTurk, campus, and community participants. *PloS One*, *11*(6), e0157732. https://doi.org/10.1371/journal.pone.0157732

Newberry, B. H. (1973). Truth telling in subjects with information about experiments: Who is being deceived? *Journal of Personality and Social Psychology*, *25*, 369–374. https://doi.org/10.1037/h0034229

Nichols, A. & Edlund, J. (2020): Why don't we care more about carelessness? Understanding the causes and consequences of careless participants, *International Journal of Social Research Methodology*, *23*, 525–638. https://doi.org/10.1080/13645579.2020.1719618

Nichols, A. L. & Maner, J. (2008). The good-subject effect: Investigating participant demand characteristics. *The Journal of General Psychology*, *135*, 151–165. https://doi.org/10:3200/GENP.1352.151-t66

Nuijten, M. B., Hartgerink, C. H., Van Assen, M. A., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*, 1205–1226. https://doi.org/10.3758/s13428-015-0664-2

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). https://doi.org/10.1126/science.aac4716

Oppenheimer, D., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*, 867–872. https://doi.org/10.1016/j.jesp.2009.03.009

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, *17*, 776–783. https://doi.org/10.1037/h0043424

Ortmann, A. & Hertwig, R. (2002). The costs of deception: Evidence from psychology. *Experimental Economics*, *5*, 111–131. https://doi.org/10.1023/A: 1020365204768

Rubin, M. (2017). When does HARKing hurt? Identifying when different types of undisclosed post hoc hypothesizing harm scientific progress. *Review of General Psychology*, *21*, 308–320. https://doi.org/10.1037/gpr0000128

Sagarin, B. J., Rhoads, K. v. L., & Cialdini, R. B. (1998). Deceiver's distrust: Denigration as a consequence of undiscovered deception. *Personality and Social Psychology Bulletin*, *24*, 1167–1176. https://doi.org/10.1177/01461672982411004

Sharpe, D. & Faye, C. (2009). A second look at debriefing practices: Madness in our method? *Ethics & Behavior*, *19*, 432–447. https://doi.org/10.1080/10508420903035455

Shimp, T. A., Hyatt, E. M., & Snyder, D. J. (1991). A critical appraisal of demand artifacts in consumer research. *The Journal of Consumer Research*, *18*, 273–283. https://doi.org/10.1086/209259

Sigall, H. & Mills, J. (1998). Measures of independent variables and mediators are useful in social psychology experiments: But are they necessary? *Personality and Social Psychology Review*, *2*, 218–226. https://doi.org/10.1207/s15327957pspr0203_5

Taylor, K. & Sheppard, J. (1996). Probing suspicion among participants in deception research. *American Psychologist*, *51*, 886–887. https://doi.org/10.1037/0003-066X.51.8.886

Tesch, F. E. (1977). Debriefing research participants: Though this be method there is madness to it. *Journal of Personality and Social Psychology*, *35*, 217–224. https://doi.org/10.1037/0022-3514.35.4.217

Walsh, W. B. & Stillman, S. M. (1974). Disclosure of deception by debriefed subjects. *Journal of Counseling Psychology*, *21*, 315–319. https://doi.org/10.1037/h0036683

Wilson, T. D., Aronson, E., & Carlsmith, K. (2010). The art of laboratory experimentation. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (eds.), *Handbook of Social Psychology*, 4th ed. (vol. 1, pp. 51–81). Wiley.

Zadvinskis, I. M. & Melnyk, B. M. (2019). Making a case for replication studies and reproducibility to strengthen evidence-based practice. *Worldviews on Evidence-Based Nursing*, *16*(1), 2–3. https://doi.org/ezproxy.library.und.edu/10.1111/wvn.12349

Zannella, L., Vahedi, Z., & Want, S. (2020). What do undergraduate students learn from participating in psychological research? *Teaching of Psychology*, *47*, 121–129. https://doi.org/10.1177/0098628320901379

PART III

Data Collection

13 Cross-Sectional Studies

Maninder Singh Setia

Abstract

Cross-sectional studies are a type of observational studies in which the researcher commonly assesses the exposure, outcome, and other variables (such as confounding variables) at the same time. They are also referred to as "*prevalence studies.*" These studies are useful in a range of disciplines across the social and behavioral sciences. The common statistical estimates from these studies are correlation values, prevalence estimates, prevalence odds ratios, and prevalence ratios. These studies can be completed relatively quickly, are relatively inexpensive to conduct, and may be used to generate new hypotheses. However, the major limitation of these studies are biases due to sampling, length-time bias, same source bias, and the inability to have a clear temporal association between exposure and outcome in many scenarios. The researcher should be careful while interpreting the measure of association from these studies, as it may not be appropriate to make causal inferences from these associations.

Keywords: Cross-Sectional Design, Biases, Statistical Methods, Advantages, Disadvantages

Introduction

A good study should be appropriately planned. Some components of a study protocol are the site of study, study design, study procedures, variables to be measured, statistical methods, and ethical aspects of the study. Thus, study design is an important aspect of any study protocol. In fact, it is often said that a badly designed study cannot be salvaged by statistical methods (Swinscow, 1997). Thus, the researcher should spend adequate time to think about the study design. Study designs can be broadly classified into two main categories: observational studies and experimental studies or intervention studies.

The two important variables in any study are the "exposure variable" and the "outcome variable." For instance, if one wants to study the effect of a particular form of therapy on schizophrenia, the therapy becomes the exposure variable and schizophrenia becomes the outcome variable. The design of the study depends on the way one handles the exposure variable. If the investigator actively modifies the exposure, then it is called an experimental study or an intervention study. For instance, in the above study, if the investigator chooses which individual will get which particular therapy for schizophrenia (psychotherapy or pharmacotherapy), the study is an intervention study – the investigator intervenes to modify the exposure. However, if the therapy has been given by some other person or is a part of the protocol, and the investigator just examined which therapy had better outcomes for management of schizophrenia, then it will be considered an observational study – the investigator does not intervene in modifying the exposure but just observes the nature of the exposure and its association with the outcome. Cross-sectional design is a type of observational study design. We will discuss various aspects of the cross-sectional study in this chapter.

Definition and Design of a Cross-Sectional Study

As discussed earlier, cross-sectional studies are a part of "observational studies." Broadly speaking, observational studies can be of three main types: cohort studies, case–control studies, and cross-sectional studies. Of course, there are other types of study designs such as case–cohort and ecologic studies. However, for the purposes of this chapter, we will restrict ourselves to these three main types of observational studies.

Cohort studies are longitudinal studies in which a group of selected participants are followed over a period, and the investigator evaluates the outcome after a certain set period of time (e.g., one year) or whenever the outcome occurs. Case–control studies have a distinctive advantage over cohort studies in terms of the "time required" to complete the study and hence they may be more efficient compared with cohort studies. In case–control studies, investigators evaluate the outcome and estimate the odds of exposure in each group. Specifically, participants who have the outcome are classified as cases and those who do not have the outcome are classified as controls.

Case–control studies were also sometimes referred to as "retrospective studies" – probably because the outcome has already occurred at the time of identification of the study participants, and the exposure is assessed after these individuals are identified. However, it may not be appropriate to term them as retrospective studies. Rather the term "prospective" and "retrospective" studies should be based on calendar time. For example, if one starts collecting data from today onward for a research study (i.e., prospectively), then the study should be termed a prospective study. However, if an investigator chooses to analyze data which are already there (collected as a part of some government survey, data from universities, or clinical data), the study is a retrospective study (prior to the calendar time). Some investigators also use the term "secondary data analyses" for studies which utilize existing data for analysis. Depending on the design and calendar time of data collection, the study may be classified as a prospective or retrospective study. These terms are applicable for case–control as well as cohort studies. Thus, we may have a prospective or retrospective cohort study, or a prospective or retrospective case–control study.

So, what is a cross-sectional study? When there is no specific structure to the sampling method or selection of study participants (either based on the exposure [cohort] or outcome [case–control]), and the information on the exposure and outcome in the study participants is collected at the same time, it is a cross-sectional study. Some authors refer to it as a "snapshot" of an underlying cohort (Szklo & Javier Nieto, 2004).

As explained earlier, depending on the calendar time of data collection, cross-sectional designs may also be classified as prospective or retrospective. I will discuss details about selection of study participants for these study designs in the subsequent paragraphs.

Participant Selection and Study Designs

Let us understand various aspects of study designs using an example. As a researcher, you have proposed to study the association between regular exercise and anxiety. The outcome is measured on a scale; the individuals can be classified into those who “have anxiety” and those who “do not have anxiety.” The definition of regular exercise is “45 minutes of exercise at least three times a week.”

Scenario 1

Let us start planning recruitment for this study. You have access to a group of individuals with varying levels of exercise. Some of them are classified in the “regular exercise group” and some are classified in the “not regular exercise group.” The level of exercise depends on the individual; you have not decided the level of exercise in these individuals. As stated earlier, since the researcher has not determined the level of exercise (i.e., exposure status), it is not an intervention study. Thus, this may be considered an observational study.

Now you have decided to recruit these individuals for your study. You recruit some individuals who are in the “regular exercise group” and some who are in the “not regular exercise group.” You assess their anxiety levels at baseline. You plan to include only those individuals who do not have anxiety at baseline. Thus, “none” of the study participants have the outcome at baseline. Now, you start following these two groups. You assess the outcome (i.e., anxiety) at three months, six months, nine months, and one year. You are interested to know what proportion of individuals have anxiety at these time points. Thus, you will measure the “incidence of anxiety” in these individuals. This is an example of a cohort study design (Figure 13.1).

In a cohort study, the researcher does not modify the exposure and the study participants are selected based on the exposure status (type of exercise in our study).

Scenario 2

Let us recruit these participants using another selection method. You have access to a group of individuals with varying levels of exercise. Some of them are classified in the “regular exercise group” and some are classified in the “not regular exercise group.” The level of exercise depends on the individual; you have not decided the level of exercise in these individuals. However, currently you also have information about their anxiety status (outcome). Some of them have anxiety and some do not. Please remember that, since the researcher has not decided the level of exercise (exposure status), it is not an intervention study. Thus, this may be considered an observational study.

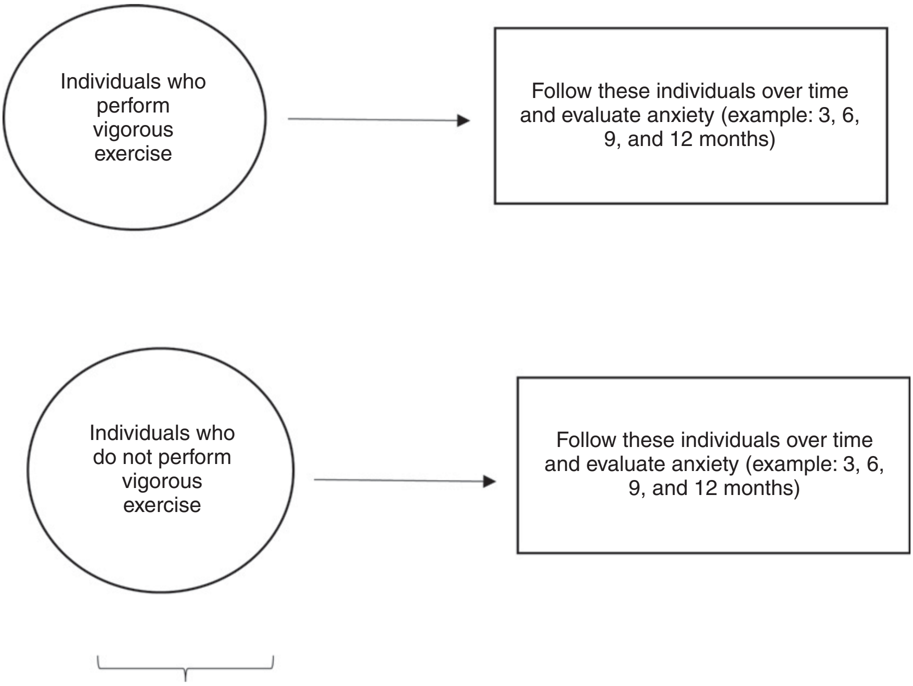

Individuals who perform vigorous exercise
Follow these individuals over time and evaluate anxiety (example: 3, 6, 9, and 12 months)
Individuals who do not perform vigorous exercise
Follow these individuals over time and evaluate anxiety (example: 3, 6, 9, and 12 months)

Figure 13.1 *Example of a cohort study.*

Now you decide to recruit the study participants using a different technique. You include some individuals who have anxiety and some who do not have anxiety. During recruitment, you are not concerned about the level of exercise. Thus, you have recruited study participants based on the outcome variable. But you still want to study the association between exercise and anxiety. So, you ask about exercise habits (how often, how long, etc. – the exposure variable) in the group with anxiety (case group) and the group without anxiety (control group). You can then classify each group in two categories. Thus, the exposure is assessed after recruiting the study participants based on the outcome. This is an example of case–control study design (Figure 13.2).

In a case–control study, the researcher cannot estimate the incidence or prevalence of the outcome. Remember, in this study you have selected the number of cases and controls (based on some statistical assumptions and calculations). Thus, the proportion of these cases (outcome is this design) is fixed by the researcher. In these studies, you can estimate the probability of exposure or the odds of exposure in each group (cases and controls). Thus, the measure of association between the exposure and the outcome is an "odds ratio."

Scenario 3

Let us consider a third scenario. The research question remains the same: What is the association between intensity of exercise and anxiety? You have access to a population in one particular area. At this point, as a researcher, you do not assign

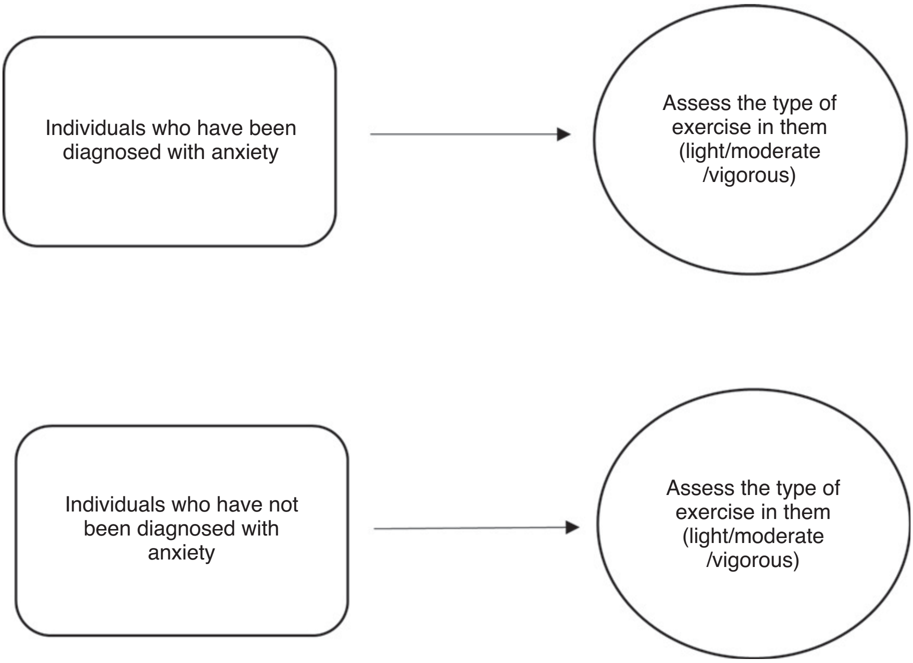

Individuals who have been diagnosed with anxiety
Assess the type of exercise in them (light/moderate /vigorous)
Individuals who have not been diagnosed with anxiety
Assess the type of exercise in them (light/moderate /vigorous)

Figure 13.2 *Example of a case–control study.*

the level/intensity of exercise to these individuals. It may sound repetitive, but the importance of this cannot be stated enough. One should not misinterpret intervention required as a part of routine care as a criterion for an intervention study. Let us understand this further. If you are taking care of an individual with schizophrenia and, as a care provider, you offer pharmacotherapy, this is an intervention. As a care provider, a surgeon may perform a surgery, and this may be called a surgical intervention. But the question you should ask yourself as a researcher is: Was this intervention done as a part of the research protocol (using some predetermined method such as randomization or other probability sampling)? If the answer is no, then this does not make it an intervention study. For an intervention study, the assignment of an intervention (i.e., exposure) should be by the investigator and for the purpose of the research study. Thus, surgical interventions, pharmacotherapeutic interventions, and psychotherapy may be considered interventions for the patient and care of the patient; however, the mere presence of these in the population under study does not make it an intervention study.

Now that we have clarified again the concept of an intervention study, let us come back to cross-sectional designs. In the population you have access to, you start recruiting participants for the study on exercise and anxiety. However, you do not recruit these participants based on their intensity of exercise (i.e., exposure variable) or presence of anxiety (i.e., outcome). The former would have made it a cohort design and the latter would have made it a case–control study. You recruit participants based on a predetermined sampling technique (such as a consecutive consenting sample or random sample). After recruiting these participants, you assess the

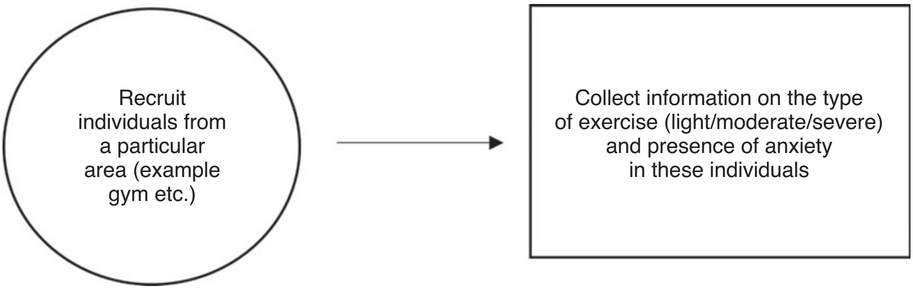

Recruit individuals from a particular area (example gym etc.)
Collect information on the type of exercise (light/moderate/severe) and presence of anxiety in these individuals

Figure 13.3 *Example of a cross-sectional study.*

intensity of exercise and presence or absence of anxiety at the same time. Thus, you have not selected these study participants based on either the exposure or the outcome and have assessed the exposure and outcome at one time point (the same time). There is no follow-up, and the study procedure for each individual ends once you have assessed the exposure, outcome, and other variables that you are interested in. This is an example of a cross-sectional study (Figure 13.3).

Features of Cross-Sectional Studies

Thus, in a cross-sectional study, the investigator collects all the information at the same time. There is no follow-up in these studies. When you assess anxiety in these individuals, you do not know whether the anxiety is a new or old occurrence of the condition. For example, in a cohort design, you only include individuals who did not have anxiety at baseline. Thus, any anxiety that was detected during follow-up of these participants was a new occurrence of anxiety. This is an estimate of the incidence of anxiety. In case–control studies, the proportion of cases is decided by the investigator (since the participants are recruited based on the presence or absence of anxiety). Thus, researchers cannot estimate either the incidence or prevalence of anxiety in the study population. As stated earlier, in cross-sectional study designs, when you identify the outcome (anxiety in this case), we do not know whether it a new occurrence or an old occurrence. Thus, this is an estimate of prevalence of the outcome (or the disease). Hence, these studies are also called "prevalence studies."

Confounding Variables

Along with the exposure and outcome, information on other variables – potential confounding factors – is also collected in these studies. What is a confounding variable? The definition of a confounding variable is: "any variable that is causally associated with the outcome, casually or non-casually associated with the exposure, and is not an intermediate variable in the casual pathway between the exposure and outcome" (Szklo & Javier Nieto, 2004; p. 180).

It is important to record information on all the potential confounding variables in any research. Knowledge of these potential confounding variables for any study is

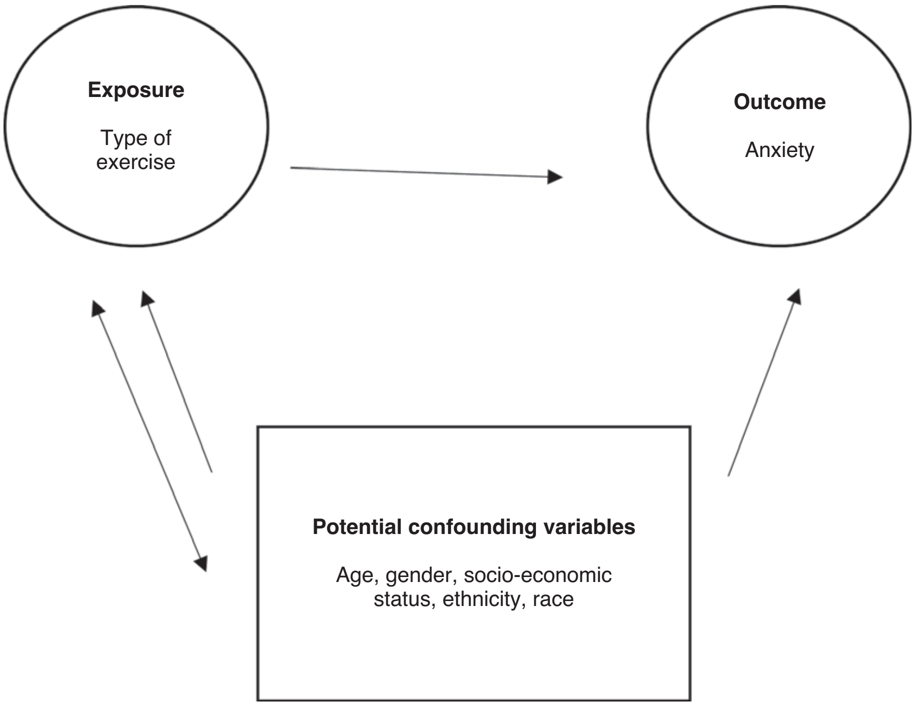

Exposure
Type of exercise
Outcome
Anxiety
Potential confounding variables
Age, gender, socio-economic status, ethnicity, race

Figure 13.4 *Example of confounding variables.*

usually based on the knowledge of literature. Thus, it is important to do a thorough literature search before designing the study; it helps a researcher to collect all information required in the study (see Chapter 4 in this volume). Some common confounding variables are age, gender, socio-economic status, ethnicity, religion, etc. Imagine that you have read in the literature that there is an association between some or all these parameters and anxiety. Thus, these are the potential confounding variables for your study and must be measured (Figure 13.4).

A simpler explanation of confounding variables is that these are variables that may help explain a part of the relationship between the exposure variable and the outcome. Let us try to understand confounding variables in the study on exercise and anxiety. Potentially, it is possible that individuals who are the lower middle or lower socio-economic status are less likely to do regular exercise. They may have limited access to gyms due to their economic conditions. Simultaneously, they may have a higher level of anxiety. This anxiety may be due to existing economic conditions; they may be anxious about paying bills or debt. If you recruit these individuals, then their exposure category will be "non-regular exercise" and the outcome category will be "anxiety." Hence, you may end up showing a relationship between exercise and anxiety – a relationship that may be considered spurious. This relationship is because of another common factor – "socio-economic status." Once one accounts for this factor in the analysis, the relationship between exercise and anxiety disappears. Thus, the interpretation of this analysis will be that "the association between exercise and anxiety is confounded by socio-economic status." Once, we accounted (or adjusted) for socio-economic status in our analysis, the relationship was not significant.

So, how do we account for/adjust for confounding variables in our studies. As a researcher, you can handle confounding variables during the design phase (such as restriction or matching) or in the analysis phase (such as by using multivariate methods). In a cohort study, if you recruit a 40-year-old male with "regular exercise," you can match and recruit a 40 (±2)-year-old male with "non-regular exercise." Thus, the comparison group is matched. In a case–control study, if you recruit a 40-year-old male with anxiety in the case group, then you can match and recruit a 40 (±2)-year-old male without anxiety in the control group. However, in a cross-sectional study design, you may not have the luxury of matching since you are assessing exposure and outcome at one point. You will have to adjust for confounding variables in the analysis phase. Hence, it is important that you collect information on all the potential confounding variables during the data collection phase.

For example, let us design another cross-sectional study:

(1) **Research question**: Is there any relation between the selection of type of course (arts, humanities, political science, technical courses, engineering, management, and medicine) and political leaning in university students?
(2) **Where will you select the participants**? Probably, the researcher will select potential participants from a university. It could be from a private university or public university. It is possible that the students in these two types of universities may be from different backgrounds. Thus, if you enroll students from only one type of university, you control for this effect. However, if you enroll students from both types of universities, you should record the type of university, and you will have to take this into account during analysis and interpretation of results.
(3) **Exposure and outcome variables**: The exposure variable for this study is the course selected and the outcome variable is political leaning (conservative, independent, liberal). For the exposure variable, you will just record the courses enrolled in and the majors of students. For the outcome variable, you propose to use a scale. The scale will generate a linear score, and the score can then be categorized as: conservative, independent, and liberal.
(4) **Potential confounding variables**: Some potential confounding variables for this study are age, gender, ethnicity, race, income, place of stay. Thus, you will collect information on these variables as well during the data collection process.
(5) **Why is this a cross-sectional study?**
    (a) Did you assign the exposure (type of course)? No. In fact, you could not have assigned the exposure status for this study (the course is chosen by the students). Thus, one should remember that it may not be possible to use all forms of designs for all research questions.
    (b) Were the students selected based on any criteria: exposure or outcome? No, you just recruited all the university students who agreed to participate in your study. Thus, this is an example of a cross-sectional study.

Where Can We Use Cross-Sectional Studies?

Population-Based Surveys

Cross-sectional designs are used for population-based surveys. These are useful for clinical, social, economic, and electoral surveys as well. A large number governmental, non-governmental, and international surveys on health (such as those by the World Bank, International Monetary Fund, Organisation for Economic Co-operation and Development, etc.) can be classified as cross-sectional studies.

The following are some examples:

(1) You are interested to know the prevalence of mental health disorders in a particular region. You decide to recruit some participants for this study using the telephone. You have decided to call the potential study participants using “random digit dialing” and administer the questionnaire for identification of mental health disorders. The estimate from this survey will be the “prevalence of mental health disorders” in the population (of that region).

(2) The researcher is interested to compare beliefs related to gender stereotypes across various countries. The researcher decides to conduct a web-based (online form) survey across various countries. The researcher can access a list of emails from 10 countries. The web-based form includes questions on demographics (age, gender, ethnicity, race, marital status, country, state/region in the country), socio-economic status (job status, monthly income), and beliefs related to gender stereotypes. This design becomes a cross-sectional study. The participants are not selected based on either the exposure or the outcome. It is a consecutive consenting sample of online participants.

Studying Serial Prevalence to Monitor Public Health Outcomes/Programs, Social Indicators, Economic Indicators, Public Administration Indicators

Cross-sectional data can be used at regular intervals to monitor the health outcomes in the population. Serial cross-sectional studies are also used to study trends in a particular population over time. In addition, they are used to monitor public policy, public administration, and economic and social indicators over time. Since it may not be possible to recruit the same individuals for monitoring these outcomes, the researchers may recruit others every year or every two years (or whatever time period has been decided). Every year, the same questionnaire is used with minor modifications. Although the same indicators are monitored over time, they are in different individuals; hence, this should be considered as a serial cross-sectional study. However, please be careful while interpreting the data from serial cross-sectional studies. The changes in the prevalence of the outcome may be an actual change or just secular changes in the various demographic and behavioral parameters over time.

The following are some examples:

(1) Sometimes in HIV programs, the prevalence is estimated in the population for a selected number of the same months (for instance, April to June) every year. This serial prevalence is then used to monitor the trends of HIV in a particular high-risk group.

(2) You are interested to understand the trends of anxiety in individuals who exercise. You choose a particular gym and recruit about 200 individuals from this gym. You visit the same gym next year and want to recruit 200 individuals to see the change in prevalence. However, this sample is different from the one recruited in the previous year. You repeat the same protocol a year later. Thus, in this study, you have conducted serial cross-sectional analysis to understand the trends of anxiety in individuals going to the gym.

(3) The researcher is interested to understand business-related behavior practices across various sectors and their changes with time. The researcher identifies certain sectors (health, hospitality, travel, defense) and asks some questions to individuals who are at least at the senior manager level (inclusion criteria for the study). You repeat the same exercise every year. Instead of recruiting the same individuals every year, you choose different individuals. This is an example of a serial cross-sectional study. As indicated earlier, the changes observed in the outcomes in business practices are due to the actual change in that sector or the change in outcomes reflects the changes in the global practices over time.

Studying the Prevalence of a Disease/Outcome and Factors Associated with It in Health-Based Settings

An example of this type of study is that you are interested in assessing the prevalence of mental health disorders in a dermatology clinic. You are attached to the clinic and recruit 250 dermatology patients (based on an estimated sample size) and evaluate the mental health disorders using a pre-validated questionnaire. You also collect information on demographics (age, gender, and socio-economic status) and want to study the association between gender and mental health disorders in these individuals.

If the study is restricted to just estimating the prevalence, some authors call is a "descriptive cross-sectional study." If your research objective is to study the association between gender and prevalence of mental health disorders it is called an "analytic cross-sectional study." Although these terms are often used in literature (Alexander et al., 2014–15), I prefer to keep it simple and simply call them "cross-sectional studies."

Generating Hypotheses That Can Be Tested Using Other Designs (Such as a Cohort Design or Trials) and Describing the Cohort at Baseline

The public health and public administration department is interested to understand the association between ethnicity of migrants and changes in beliefs about gender stereotypes over time. They have decided to use a cohort study design (since they have the resources to follow a large number of individuals over time). They recruit individuals from different cities and assess their beliefs about gender stereotypes every year. After an initial analysis at baseline, there appears to be a relation between socio-economic status and beliefs about gender stereotypes as well as a relation between the country of origin (migrants) and gender stereotypes. Thus, two new research questions can be generated from these baseline analyses, even though the study design is a cohort design. These hypotheses can be tested using other cross-sectional studies or cohort studies.

Diagnostic Test Property/Diagnostic Accuracy Studies

Diagnostic test property and accuracy studies examine the properties of a new diagnostic test. This could be a laboratory-based test, a clinical algorithm, or a new psychological scale. The usual design in these studies is cross-sectional. The researcher selects a sample (some have the outcome/disease, and some do not). Everyone is administered both the tests – the one that the researcher wants to assess and the existing gold standard of diagnosis. The researcher compares the results from both these tests conducted on the same subject/same tissue sample. Diagnostic test properties such as sensitivity, specificity, and positive and negative predictive values are estimated from these studies.

Biases in a Cross-Sectional Study

The biases in this type of study design may be due to the design of the study or due to the sampling frame and methods used.

Length-Time Bias

Length-time bias is of particular concern in cross-sectional studies. It is important to remember that, in cross-sectional studies, information on the exposure and outcome is collected only once at the same time. Thus, when we recruit these participants, the outcome should be present when we assess the study participant. This is more common in health outcomes such as cancers or other chronic conditions. If the individual has a milder form of the disease, it is likely that many of these individuals may not show the outcome when you assess them cross-sectionally (at one point of time). If you were to follow the same individuals after some time, you may find that outcome is present in these individuals, but it was not detected during the time of the study. Another possibility is that, if the disease is very severe and the fatality is high and early, individuals who have this form of the disease will not survive long enough to be a part of this “snapshot.”

Consider the following example: When antiretroviral treatment was not universally available, the slow progressors of HIV were over-represented in a prevalence study compared with rapid progressors who died early after acquiring the infection. It is quite likely that the immunologic and clinical features in slow progressors may

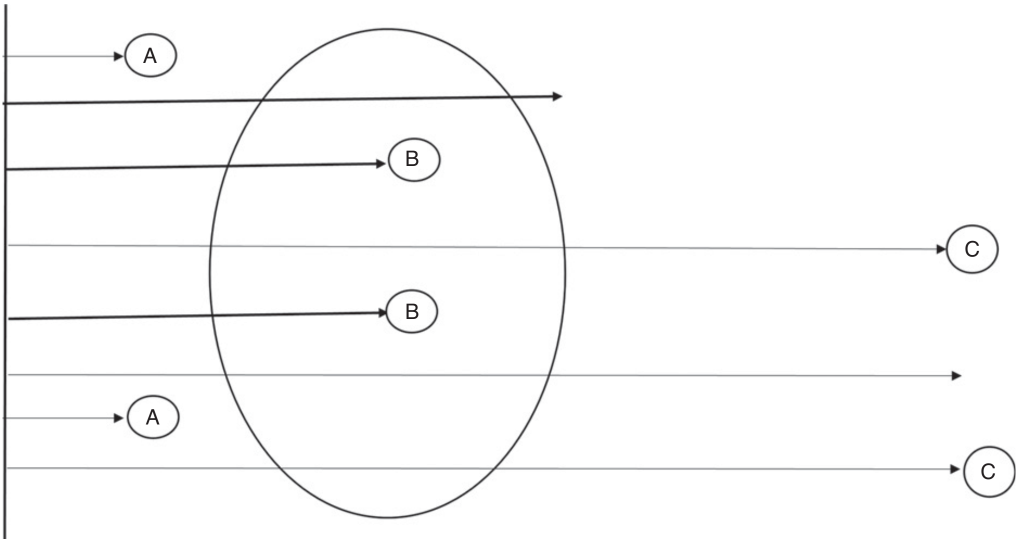


Figure 13.5 *An example of length-time bias.*

be different from rapid progressors. Thus, the estimates and associations from this cross-sectional study may be biased.

Let us understand this by considering Figure 13.5. In these individuals, the initiation of the disease occurs at the same time (vertical straight line). The circle is the time and extent of your cross-sectional study. On one hand, for some of these cases, the fatality is very high (group A). Thus, they do not even reach the circle (when you have conducted the study). One the other hand, in some of these cases (group C), the disease is very mild and only manifests after the duration of the cross-sectional study. Thus, out of these six cases, the researcher has only identified two cases (group B) for that cross section of time.

Temporality

There may be biases because of temporality issues in cross-sectional studies. Since the exposure and the outcome status is collected at the same time, the temporal relation and causation between these two often cannot be established. In general, one is not able to comment on whether the exposure came first or the outcome. This is often a concern for social and behavioral science research. Certain behaviors may change after the occurrence/knowledge of the outcome. For instance, food and smoking behaviors may change after knowledge of having certain diseases (such as diabetes, lung diseases, or cardiovascular diseases). The correlations and associations usually measured in cross-sectional studies should not be confused with causation.

For example, you plan to conduct a study to assess the relation between physical activity and body mass index (BMI). You design a cross-sectional study of 200 participants. You collect information on the intensity of physical activity (light/moderate/vigorous etc.) and the current BMI.

Before interpreting the association, you may have to account for the temporal sequence of events. It is quite likely that those with high BMIs have started to engage in vigorous physical activity (exercise) after the knowledge of their BMI. Thus, when you do a cross-sectional analysis, you may find that a higher proportion of individuals who do vigorous exercise also have high BMI!

So how can we handle this temporal bias?

- By doing a cohort/longitudinal study, the temporality of events can be recorded. For example, in the above study, if you assess the exercise type and BMI, you can record when the individual started exercising more (before or after the increase in BMI). Of course, this may not be feasible always. Some of the variables (e.g., birth history) may have a clear temporal structure in many studies.
- Design the questionnaire to get as much information on the occurrence of outcome (if known), record changes in behaviors, and measure other potential confounding variables
- Some outcomes that have occurred in the specific time period may be excluded. For example, Szklo and Nieto (2004) have highlighted that epidemiologists may exclude deaths that have been recorded in a certain time of the study.

Same-Source Bias

This is an important bias in cross-sectional studies, particularly in social and behavioral research. Many behaviors overlap, and self-reported exposure and outcome variables in the same individual may have errors. These errors in one measurement may also be correlated with errors in the measurement of another variable if it is measured in the same individual. Thus, any correlation or association between these self-reported variables (exposure and outcome) may be found when it does not exist.

Faveo and Bullock (2015) define common method bias as, "a biasing of results (which could be in the form of false positives from hypothesis tests) that is caused by two variables exhibiting related measurement error owing to a common method, such as a single survey" (Faveo and Bullock, 2015, p. 285). The authors have used this definition for public administration; they suggest that since most of the variables are based on surveys and perception, the association in these studies may be affected by common-source bias. They have also suggested many methods to handle this issue such as: Harman's single-factor test, Brewer's split sample method, marker variables, and differencing. Furthermore, George and Pandey (2017) have presented a useful flow chart to address common source bias for studies in public administration. A detailed discussion of these methods is beyond the scope of this chapter, and I strongly encourage you to read the above-mentioned references to understand these techniques.

A very good example of same-source bias is in a letter by Gullon and colleagues (Gullon et al., 2014), written in response to a study on urban environment and physical activity (Rodriguez-Romo et al., 2013). This study had found a relationship between the attributes of the neighborhood and physical activity. However, Gullon et al. (2014) argued that since both the outcome (physical activity) and exposure (neighborhood) were self-reported, it is likely that individuals who are less active may perceive their neighborhood not favorable for physical activity. Thus, according to them, this relationship could be biased because of same-source data collection. One method of reducing this bias would be to use actual environmental assessments (physical verification) rather than self-reporting.

Biases in Sampling

There can be additional biases due to sampling and response. For example, in the above-mentioned population-based survey (anxiety and exercise), if you choose the sample from a given population using the random digit dialing, it is likely that you may miss the population with low phone coverage (typically low socio-economic status). Furthermore, if the questionnaire for anxiety is available only in English, many people will not be eligible for inclusion in the study (i.e., if they do not speak English – this may also be related to immigration and socio-economic status). Thus, there may be a *selection bias* while enrolling the study participants. If a certain group of individuals are more likely to not respond to your calls or not agree to participate, then the estimates from the study may be biased. For example, it is possible that individuals who are aware of their anxiety may not respond to your call – the *non-response bias*.

In surveys, individuals may sometimes be wary of giving responses that go against the main narrative. They may give responses that they think the interviewer would want to hear, rather their true responses. This is usually seen in election surveys when the respondents usually pick a choice for which they will not be judged. Similarly, in clinical or psychological settings, they may over-report behaviors which may be considered positive in the community (e.g., use of condoms) and under-report behaviors which may be considered negative in the community (e.g., use of drugs or cigarettes). This may lead to *social desirability bias* (see Chapter 11 in this volume).

In clinic-based sampling, only those who access the clinic will be eligible for inclusion in the study. These individuals may have a higher awareness of the condition or may have a more positive health-seeking behavior (the fact they have approached a health care facility) compared with those in the community. It is also quite likely that these individuals may have a severe form of the disease. Thus, all these aspects must be considered while interpreting the findings from a clinic-based cross-sectional study.

For example, let us use the same study: Is there any relation between selection of the type of course (arts, humanities, political science, technical courses, engineering, management, and medicine) and political leaning in university students? What are the potential sources of bias in this study?

- **Selection bias:** Hypothetically, if the students in arts and humanities who lean toward being conservative are concerned that they will be judged due to their responses do not consent to participate in the study (due to the belief that students in these courses are of liberal leaning), you may have an under-representation of these students. Thus, you will end up overestimating the relationship between these courses and a liberal political leaning.
- **Temporal bias:** You enroll students from these courses; these students are in various years of their program (first/second/third/fourth). Some of these students may have changed their political leanings after their entry into the course due to the course material, peer pressure, or for any other reason.

Thus, the temporal nature of the association cannot be ascertained in such a scenario after statistical correlation or association. There are two questions: Is there a relation between choice of course and political leaning? or Does entry into a course have any effect on the political leaning in university students?

So, how can we address this?

- If you include only students in their first year – probably in the first three months after joining the course – it is less likely that political leanings may have changed so soon after joining the course. Thus, the estimates may be less biased.
- Interestingly, if you do the same study every year among first-year students in the first three months, this becomes a serial cross-sectional study. In this study, you will be able to study the changes in pattern of the relation/association over time.

Analysis of Cross-Sectional Data

The common measures of outcome and association estimated in a cross-sectional study are correlation, prevalence, prevalence ratios, and odds ratios.

Correlation

A large proportion of social and behavioral science research is correlational. Broadly speaking, a correlation estimates the linear relationship between two variables (e.g., the exposure and outcome). Depending on the nature of the variables, the correlation can be estimated by different methods.

The relationship between two linear variables that are normally distributed is estimated by Pearson's correlation coefficient and is denoted as *r*. This correlation coefficient measures the linear relationship between these two variables. The relationship between non-parametric linear variables or ordinal variables can be estimated using the Spearman's rank correlation coefficient, and it denoted as ρ (rho). This measures the monotonic relationship between two variables.

A scatter plot can be used to check the relationship between these two variables (see Figure 13.6, in next section). The values of correlation can range from −1 to +1 (includes 0). A value of +1 or −1 indicates perfect correlation; the former indicates a perfect positive correlation (the outcome variable increases with an increase in the exposure variable), and the latter indicates a perfect negative correlation (the outcome variable decreases with an increase in the exposure variable). Thus, a correlation value closer to one indicates a strong correlation and the sign of the estimate (positive or negative) indicates the direction of the correlation (see the detailed discussion in the next section).

Similarly, a correlation value of zero indicates no correlation; hence, a value close to zero will be considered a weak correlation.

Prevalence

A simple definition of the prevalence is the “proportion of outcome (old and new) in the sample.”

For example, you have included 200 (*N*) individuals from the gym in your study, and you report that 60 (*n*) of them have anxiety. At this point, you do not know whether the condition is old (i.e., present for a long time) or new. Here, the prevalence = *n*/*N* or 60/200 = 30%. Some authors refer to this prevalence as the “point prevalence” since this gives us a prevalence at one point of time.

Odds Ratios

Odds ratios (ORs) are a common measure of association in epidemiological studies. In cross-sectional studies, these odds ratios are also called prevalence odds ratios (PORs).

Example: You wish to study the association between gender and anxiety in the above study. Let us construct the 2 × 2 table (see Table 13.1):

- The odds ratio = ratio of odds of the outcome in the exposed group to the odds in the unexposed group. The odds of outcome (anxiety) in the exposed group (females in our study): a/b.
- The odds of outcome (anxiety) in the unexposed group (males in our study): c/d.
- The ratio of these two odds? $(a/b)/(c/d)$.

Thus:

$$\begin{aligned} \text{POR} &= ad/bc \\ &= (30 \times 50)/(30 \times 90) \\ &= 0.56. \end{aligned}$$

Since it is less than 1, exposure is protective (i.e., the odds of having anxiety if one is a female is 0.56 compared with a male in individuals coming to the gym).

Prevalence Ratio

In cross-sectional studies, one can also estimate the prevalence ratio (PR). For example, let us use the same study:

$$\begin{aligned} \text{Prevalence of anxiety in females} &= a/(a+b) \\ &= 30/120 \end{aligned}$$

Table 13.1 *The association between gender and anxiety*

	Anxiety – Yes	Anxiety – No	Total
Females	30 (*a*)	90 (*b*)	120
Males	30 (c)	50 (*d*)	80
Total	60	40	200

Prevalence of anxiety in males = $c/(c + d)$
= 30/80
PR = (30/120)/(30/80) = 0.67

So, which estimate do you use? Let us consider multiple scenarios in this same example (Table 13.2).

What do we observe here? In general, the POR overestimated the association (the strength of association was stronger in the POR compared with the PR). When the overall prevalence of the outcome is low (<5%), the POR is similar to PR. Some authors recommend that the POR should be used for cross-sectional studies with chronic conditions and PR should be used for acute conditions (Alexander et al., 2014–15). However, others have argued that the PR may be a more appropriate measure of association as it has better interpretability compared with the POR (Lee & Chia, 1994).

There are numerous articles which support or oppose the use of either of these measures (PR vs POR) in cross-sectional studies or prevalence data. As a researcher:

- It is best that you pick either the PR or POR as a measure of your choice.
- You should justify it adequately by using references from the literature (some of these are provided in the references) including some published studies that have used your measure.
- The measures should be appropriately estimates and interpreted.
- After selecting the measure of association (PR or POR), use the corresponding multivariate model (discussed in the next section).

Multivariate Analyses

Multivariate analyses are used for adjusting for potential confounding variables in a study. Though many researchers use logistic regression models for multivariate analysis in cross-sectional studies, the problems with such models have been acknowledged. Indeed, some authors (Lee et al., 2009) have discouraged the use of these models since, as seen above, the ORs may not be appropriate approximations of PRs. Thus, alternative models that are good estimates of the PR, such as Cox regression with robust variance, Poisson regression with robust variance, and log-binomial regressions, are more appropriate for cross-sectional studies. However, it has also been suggested that logistic regression with random effects models may be used for cluster cross-sectional studies.

Table 13.2 *The multiple scenarios of the study of association between gender and anxiety*

	Anxiety – Yes	Anxiety – No	Total	Prevalence of outcome (*n*3/*N*)	POR *ad*/*bc*	PR (*a*/*n*1)/(*c*/*n*2)
Scenario 1						
Females	30 (*a*)	90 (*b*)	120 (*n*1)	30%	0.56	0.67
Males	30 (*c*)	50 (*d*)	80 (*n*2)			
Total	60 (*n*3)	140 (*n*4)	200 (*N*)			
Scenario 2						
Females	20	100	120	20%	0.60	0.67
Males	20	60	80			
Total	40	160	200			
Scenario 3						
Females	15	105	120	15%	0.62	0.67
Males	15	65	80			
Total	30	170	200			
Scenario 4						
Females	10	110	120	10%	0.64	0.67
Males	10	70	80			
Total	20	180	200			
Scenario 5						
Females	5	115	120	5%	0.65	0.67
Males	5	75	80			
Total	10	190	200			
Scenario 6						
Females	2	118	120	2%	0.66	0.67
Males	2	78	80			
Total	4	196	200			

It is beyond the scope of this chapter to discuss all the models in detail. Hence, I refer the reader to other relevant chapters in this volume and to the references that have been provided at the end of this chapter.

As an example. let us analyze the same study: Is there any relation between selection of the type of course (arts, humanities, political science, technical courses, engineering, management, and medicine) and political leaning in university students? Apart from the main exposure (type of course) and outcome (political leaning), you are interested to study the relation between age (potential confounding variable) and political leaning. You have measured the age in years and political leaning is also on a (hypothetical) linear scale of 0–20 (scores close to 20 indicate conservative leaning and those close to zero indicate liberal leaning). These are two linear variables. Thus, you plan to use Pearson's correlation coefficient *r* to study the relation between these two variables (Figure 13.6).

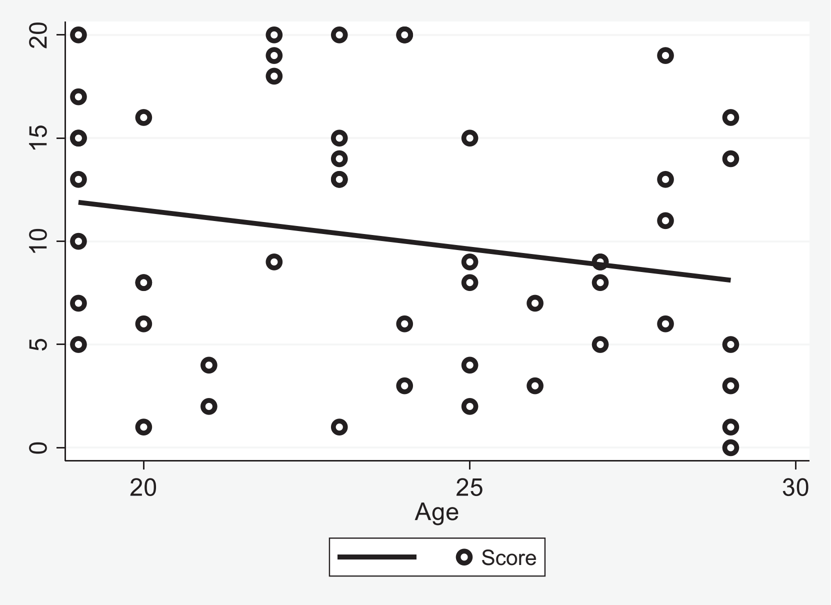


Figure 13.6 *Scatter plot and correlation between age and scores of political leaning in a hypothetical population.*

As seen in the scatter plot, age is plotted on the *x*-axis and score is plotted on the *y*-axis. The solid line is the fitted line, and the linear trend shows a declining trend (i.e., as age increases the scores decrease). The estimated Pearson's correlation value *r* is −0.21. It is closer to "0" compared with "1". However, it may be considered a relatively good correlation in some fields (e.g., psychology). The negative sign of the correlation fits with the declining linear trend seen in the scatter plot. The *p*-value was 0.14; it appears to be non-significant.

Advantages and Disadvantages of Cross-Sectional Studies

The *advantages* of cross-sectional studies are as follows:

- Due to the nature of the design, these studies are relatively easy to conduct and can be usually completed within a short span of time. The researcher recruits the study participants and collects the information on the exposure variable, outcome variable, and confounding variables at the same time. The most important thing is to read the literature thoroughly so that you prepare the list of variables to be collected.
- A cross-sectional study is a snapshot of exposure and outcome variables (and confounding variables) at the same time. Thus, we collect information of multiple variables at the same time. The association between other variables (other than the primary exposure and outcome variable) can also be estimated in cross-sectional studies. However, one should not make this a "fishing exercise."

- These studies can be used to generate hypotheses that can be further tested using other study designs (such as cohort design or randomized controlled trials).
- These studies are useful to monitor health outcomes/service delivery indicators in health programs.
- The studies are useful (and sometimes the only option) for studies in social and behavioral science disciplines.

However, the *disadvantages* of cross-sectional studies are:

- As discussed earlier, it may be difficult to ascertain the temporal relationship between the exposure and outcome in the cross-sectional design.
- The interpretation of results from these studies should be within the context of various biases (as discussed above). A correlation between variables should not be considered causation.
- It may be difficult to do cross-sectional studies of diseases of short duration or those with an extremely high fatality rate.

Conclusion

Though there may be multiple limitations of this design, cross-sectional studies are useful for timely data collection. This design may be used in different specialties and for different types of outcomes, such as correlation, prevalence, association between multiple exposures and outcomes, health service/delivery, public administration, political science (democracy/populism/political leanings), sociological outcomes (gender/ethnicity/race/culture/poverty related studies), psychological studies (peer pressure/depression/mental health/memory/sleep), and diagnostic accuracy studies. However, the researcher should be careful when interpreting the results of these studies; it may not be appropriate to make causal inferences from these associations.

Further Reading

The following are sources that describe various aspects of cross-sectional studies.

Axelson, O., Fredriksson, M., & Ekberg, K. (1994). Use of the prevalence ratio v the prevalence odds ratio as a measure of risk in cross sectional studies. *Occupational and Environmental Medicine*, *51*(8), 574. https://doi.org/10.1136/oem.51.8.574

Barros, A. J. & Hirakata, V. N. (2003). Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. *BMC Medical Research Methodology*, *3*, 21. https://doi.org/10.1186/1471-2288-3-21

Brumback, B. & Berg, A. (2008). On effect-measure modification: Relationships among changes in the relative risk, odds ratio, and risk difference. *Statistics in Medicine*, *27*(18), 3453–3465. https://doi.org/10.1002/sim.3246

Campbell, D. & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105.

Colditz, G. A. (2010). Overview of the epidemiology methods and applications: strengths and limitations of observational study designs. *Critical Reviews in Food Science and Nutrition*, *50 Suppl 1*, 10–12. https://doi.org/10.1080/10408398.2010.526838

Freemantle, N., Marston, L., Walters, K., et al. (2013). Making inferences on treatment effects from real world data: propensity scores, confounding by indication, and other perils for the unwary in observational research. *BMJ*, *347*, f6409. https://doi.org/10.1136/bmj.f6409

Garger, J. (2020). A definition of single source bias in social science research. Available at: www.johngarger.com/blog/a-definition-of-single-source-bias-in-social-science-research.

Hennekens, C. H. & Buring, J. E. (1987). *Epidemiology in Medicine*. Lippincott Williams & Wilkins.

Hughes, K. (1995). Odds ratios in cross-sectional studies. *International Journal of Epidemiology*, *24*(2), 463–464, 468. https://doi.org/10.1093/ije/24.2.463

Hulley, S, B., Cummings, S. R. Browner, W. S., et al. (2001). *Designing Clinical Research,* 2nd ed. Lippincot Williams & Wilkins.

Jewell, N. (2004). *Statistics for Epidemiology.* Chapman and Hall/CRC.

Kleinbaum, D., Kupper, L., & Morgenstern, H. (1982). *Epidemiologic Research*. John Wiley & Sons.

Lee, J. (1994). Odds ratio or relative risk for cross-sectional data? *International Journal of Epidemiology*, *23*(1), 201–203. https://doi.org/10.1093/ije/23.1.201

Martinez, B. A. F., Leotti, V. B., Silva, G. S. E., et al. (2017). Odds ratio or prevalence ratio? An overview of reported statistical methods and appropriateness of interpretations in cross-sectional studies with dichotomous outcomes in veterinary medicine. *Frontiers in Veterinary Science*, *4*, 193. https://doi.org/10.3389/fvets.2017.00193

Mellis, C. M. (2020). How to choose your study design. *Journal of Paediatrics and Child Health*, *56*(7), 1018–1022. https://doi.org/10.1111/jpc.14929

Mitchell, T. (1985). An evaluation of the validity of correlational research conducted in organizations. *Academy of Management Review*, *10*(2), 192–205.

Moore, D., & McCabe, G. (2002). *Introduction to the Practice of Statistics,* 4th ed. WH Freeman and Company.

Mukaka, M. M. (2012). Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, *24*(3), 69–71. https://www.ncbi.nlm.nih.gov/pubmed/23638278

Pandis, N. (2014a). Cross-sectional studies. *American Journal of Orthodontics and Dentofacial Orthopedics*, *146*(1), 127–129. https://doi.org/10.1016/j.ajodo.2014.05.005

Pandis, N. (2014b). Introduction to observational studies: part 2. *American Journal of Orthodontics and Dentofacial Orthopedics*, *145*(2), 268–269. https://doi.org/10.1016/j.ajodo.2013.11.002

Pearce, N. (2004). Effect measures in prevalence studies. *Environmental Health Perspectives*, *112*(10), 1047–1050. https://doi.org/10.1289/ehp.6927

Polychronopoulou, A., & Pandis, N. (2014). Interpretation of observational studies. *American Journal of Orthodontics and Dentofacial Orthopedics*, *146*(6), 815–817. https://doi.org/10.1016/j.ajodo.2014.10.004

Reichenheim, M. E. & Coutinho, E. S. (2010). Measures and models for causal inference in cross-sectional studies: arguments for the appropriateness of the prevalence odds ratio and related logistic regression. *BMC Medical Research Methodology*, *10*, 66. https://doi.org/10.1186/1471-2288-10-66

Rothman,K. J. Greenland, S., & Lash, T. L. (2008). *Modern Epidemiology,* 3rd ed. Lippincott Williams & Wilkins.

Santos, C. A., Fiaccone, R. L., Oliveira, N. F., et al. (2008). Estimating adjusted prevalence ratio in clustered cross-sectional epidemiological data. *BMC Medical Research Methodology*, *8*, 80. https://doi.org/10.1186/1471-2288-8-80

Sedgwick, P. (2014). Spearman's rank correlation coefficient. *BMJ*, *349*, g7327. https://doi.org/10.1136/bmj.g7327

Setia, M. S. (2016). Methodology series module 3: Cross-sectional studies. *Indian Journal of Dermatology*, *61*(3), 261–264. https://doi.org/10.4103/0019-5154.182410

Sedgewick, P. (2018). Spearman's rank correlation coefficient, (correction). *BMJ*, *362*, k4131. https://doi.org/10.1136/bmj.k4131

Stromberg, U. (1994). Prevalence odds ratio v prevalence ratio. *Occupational and Environmental Medicine*, *51*(2), 143–144. https://doi.org/10.1136/oem.51.2.143

Tamhane, A. R., Westfall, A. O., Burkholder, G. A., & Cutter, G. R. (2017). Prevalence odds ratio versus prevalence ratio: choice comes with consequences. *Statistics in Medicine*, *36*(23), 3760. https://doi.org/10.1002/sim.7375

Thiese, M. S. (2014). Observational and interventional study design types; an overview. *Biochemia Medica*, *24*(2), 199–210. https://doi.org/10.11613/BM.2014.022

Thompson, M. L., Myers, J. E., & Kriebel, D. (1998). Prevalence odds ratio or prevalence ratio in the analysis of cross sectional data: what is to be done? *Occupational and Environmental Medicine*, *55*(4), 272–277. https://doi.org/10.1136/oem.55.4.272

Thorndike, E. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, *4*(1), 25–29.

Traissac, P., Martin-Prevel, Y., Delpeuch, F., & Maire, B. (1999). Regression logistique vs autres modeles lineaires generalises pour l'estimation de rapports de prevalences. [Logistic regression vs other generalized linear models to estimate prevalence rate ratios.] *La Revue d'épidémiologie et de santé publique 47*(6), 593–604. https://www.ncbi.nlm.nih.gov/pubmed/10673593

Twisk, J. W. R.(2013). *Applied Longitudinal Data Analysis for Epidemiology,* 2nd ed. Cambridge University Press.

Zocchetti, C., Consonni, D., & Bertazzi, P. A. (1995). Estimation of prevalence rate ratios from cross-sectional data. *International Journal of Epidemiology*, *24*(5), 1064–1067. https://doi.org/10.1093/ije/24.5.1064

References

Alexander, L., Lopes, B., Ricchetti-Masterson, K., & Yeatts, K. (2014–15). Cross-sectional Studies. UNC CH Department of Epidemiology. Available at: https://sph.unc.edu/wp-content/uploads/sites/112/2015/07/nciph_ERIC8.pdf.

Favero, N. & Bullock, J. (2015). How (not) to solve the problem: An evaluation of scholarly responses to common source bias. *Journal of Public Administration Research and Theory*, *25*(1), 285–308.

George, B. & Pandey, S. K. (2017). We know the yin – but where is the yang? Toward a balanced approach on common source bias in public administration scholarship. *Review of Public Personnel Administration*, *37*(2), 245–270. https://doi.org/10.1177/0734371X17698189

Gullon, P., Bilal, U., & Franco, M. (2014). Physical activity environment measurement and same source bias. *Gaceta Sanitaria*, *28*(4), 344–345. https://doi.org/10.1016/j.gaceta.2013.12.011

Lee, J. & Chia, K. S. (1994). Use of the prevalence ratio v the prevalence odds ratio as a measure of risk in cross sectional studies. *Occupational and Environmental Medicine*, *51*(12), 841. https://doi.org/10.1136/oem.51.12.841

Lee, J., Tan, C. S., & Chia, K. S. (2009). A practical guide for multivariate analysis of dichotomous outcomes. *Annals of the Academy of Medicine, Singapore*, *38*(8), 714–719. https://www.ncbi.nlm.nih.gov/pubmed/19736577

Rodriguez-Romo, G., Garrido-Munoz, M., Lucia, A., Mayorga, J. I., & Ruiz, J. R. (2013). Asociacion entre las caracteristicas del entorno de residencia y la actividad fisica. [Association between the characteristics of the neighborhood environment and physical activity.] *Gaceta Sanitaria*, *27*(6), 487–493. https://doi.org/10.1016/j.gaceta.2013.01.006

Swinscow, T. (1997). Study design and choosing a statistical test. BMJ Publishing Group. Available at: www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/13-study-design-and-choosing-statisti.

Szklo, M. & Javier Nieto, F. (2004). *Epidemiology: Beyond the Basics*. Jones and Bartlett Learning.

14 Quasi-Experimental Research

Charles S. Reichardt, Daniel Storage, and Damon Abraham

Abstract

In this chapter, we discuss the logic and practice of quasi-experimentation. Specifically, we describe four quasi-experimental designs – one-group pretest–posttest designs, non-equivalent group designs, regression discontinuity designs, and interrupted time-series designs – and their statistical analyses in detail. Both simple quasi-experimental designs and embellishments of these simple designs are presented. Potential threats to internal validity are illustrated along with means of addressing their potentially biasing effects so that these effects can be minimized. In contrast to quasi-experiments, randomized experiments are often thought to be the gold standard when estimating the effects of treatment interventions. However, circumstances frequently arise where quasi-experiments can usefully supplement randomized experiments or when quasi-experiments can fruitfully be used in place of randomized experiments. Researchers need to appreciate the relative strengths and weaknesses of the various quasi-experiments so they can choose among pre-specified designs or craft their own unique quasi-experiments.

Keywords: Quasi-Experiments, Research Design, Threats to Internal Validity, Pretest–Posttest Design, Nonequivalent Group Design, Regression Discontinuity Design, Interrupted Time-Series Design

Introduction

Quasi-experiments and randomized experiments are both used to assess the effects of treatments, programs, and interventions. A quasi-experimental design is like a randomized experiment in that a comparison is drawn between (1) outcomes following a treatment and (2) outcomes following an alternative treatment, which might be no treatment at all (i.e., a control group). The difference between the two types of designs lies in how the treatment and alternative treatment conditions are assigned. In randomized experiments, the treatment and alternative treatment conditions are assigned at random. In quasi-experiments, treatment conditions are not assigned at random, and therein lies a difficulty that must be confronted in quasi-experimentation, as will be explained as we proceed.

Randomized experiments are often considered the gold standard of research designs for estimating treatment effects. However, quasi-experiments can often be adequate substitutes for randomized experiments and can even be preferred when research circumstances are not conducive to randomized experiments. For example, either practical or ethical constraints can make the random assignment of treatments unacceptable. Under such circumstances, researchers are forced to rely on quasi-experiments. Fortunately, researchers can often do so to good effect (Reichardt, 2019; Shadish et al., 2002). Understanding the workings of quasi-experimental designs requires first recognizing the role played by threats to internal validity. Given that preliminary recognition, we then explicate the logic of four prototypical quasi-experimental designs.

Threats to Internal Validity

Estimating a treatment effect requires comparing outcomes following the receipt of a treatment, on the one hand, to outcomes following the receipt of no treatment or an alternative treatment, on the other hand. Although our discussion focuses on comparing only two treatment conditions, our conclusions generalize to comparisons between more than two treatment conditions. Also, note that the participants in a comparison could be either individual people or groups of people (e.g., classrooms of students or even whole cities or nations). In addition, treatment conditions could be imposed either by people or by nature, such as when assessing the effects of sex differences.

Unfortunately, it is not possible to estimate a treatment effect by drawing a comparison between treatment conditions, without varying something else besides the different treatments. That is, a researcher cannot implement both treatment X and an alternative treatment Y instead, with everything else being the same. Of course, a treatment effect might be estimated by comparing what happens after giving one group treatment X and what happens after giving a different group treatment Y instead. But then the people in the two conditions vary along with the treatments received, so everything else besides the treatments would not be the same. In any practical comparison between treatment conditions, something else besides the treatments must differ across the treatment conditions.

Whatever differs across the treatment conditions in a comparison is confounded with the treatment conditions. Confounding conditions are also called threats to internal validity. Because something else will differ across treatment conditions, in addition to the treatments received, at least one threat to internal validity will always arise when estimating treatment effects. The problem is that threats to validity can bias estimates of treatment effects. To avoid bias in estimating treatment effects, the effects of threats to internal validity must be considered. The different threats to internal validity that are present in different quasi-experiments and ways to take their effects into account will be described as we introduce the four prototypical quasi-experiments next.

One-Group Pretest–Posttest Designs

The one-group pretest–posttest design is one of the simplest quasi-experimental designs. The prototypical one-group pretest–posttest design has the following structure: A pretest is assessed on a single group of participants, a treatment is introduced, and a posttest is then assessed on the same group of participants. The results from the pretest are compared to the results on the posttest, and the difference in outcomes is used to estimate the effects of the treatment.

Examples of the One-Group Pretest–Posttest Design

The one-group pretest–posttest design has a long history in social and behavioral research. For example, Eysenck (1952) reported the results of an extensive collection of one-group pretest–posttest studies investigating the effects of psychotherapy on psychological well-being. The results revealed that a substantial proportion of clients improved from before to after treatment; this was attributed to the effects of the intervening treatment, though (as we comment in a later section) Eysenck disputed that interpretation, using a more elaborate quasi-experimental design.

The one-group pretest–posttest design is also widely used in studies of the effects of educational interventions. For example, Arum and Roksa (2010) assessed the intellectual abilities of college students during their first term in school and at the end of their second year, with two years of college intervening in between. Unfortunately, only a small difference between these two measurements was observed, which Arum and Roksa attributed to the relatively small effect of a typical college education on cognitive skills (though that interpretation has not been without its critics). In another educational example, St. Pierre et al. (1999) reported that the one-group pretest–posttest design was the predominant design used in hundreds of local evaluations across the United States to assess the effectiveness of the Even Start program for improving family literacy.

Threats to Internal Validity

A variety of threats to the internal validity in the one-group pretest–posttest design can weaken an inference about a treatment effect. We describe eight potential threats to internal validity in the one-group pretest–posttest design in the context of a hypothetical example – a program intended to reduce depression in students during their first year in college; depression is assessed by self-reports of the students, both before and after participation in the program.

First, there is a threat due to *history*, meaning some external event besides the treatment took place after the pretest but before the posttest, which had an effect on the posttest and biased the posttest assessment. For example, the students in the program may have also been enrolled in other programs to ease the transition to college and these other programs produced changes in depression between pretest and posttest.

Second, there is a threat due to *maturation* because participants grow older between the time of the pretest and posttest in a way that might bias the estimate of treatment effects. For example, there might be a natural progression in which first-year students become less depressed as they become acclimated to being in college and away from home, even without the program to reduce depression. As a result, any differences from pretest to posttest might have been because of the natural progression rather than the program.

The third threat, due to *regression toward the mean*, arises when the pretest is collected at a time where the participants' outcomes are either better or worse than average and, as a result, are expected to revert to more typical levels of performances by the time of the posttest. For example, if the program were offered to volunteers, students might have chosen to enroll in the program precisely because they were suffering from unusually high levels of depression; their depression might have improved on its own by the time of the posttest, simply because average levels of mental health are more common than unusually high levels.

A fourth threat to internal validity is *testing*, whereby the mere collection of the pretest influences outcomes at the time of the posttest. For example, being asked their degree of depression on a pretest might have made the students aware of just how serious their depression was and inspired them to alter their level of depression on their own, even if they had not participated in the program.

Fifth, a threat of *instrumentation* arises when the measurement instrument (in the example, the students themselves because they were making self-reports) changes from pretest to posttest, in the absence of real change in the underlying conditions. For example, the students may have recalibrated their assessment of the degree of their depression that led to a change on the posttest, even when there were no actual changes in the level of depression from pretest to posttest.

A sixth threat is due to *cyclical changes* (also called *seasonality*), where the pretest and posttest were collected at different times during a regular cycle (e.g., at different times of day, days of the week, and seasons of the year) and the differences between the times in the cycles account for differences in the participants' performances from pretest to posttest. For example, if the program was administered during the fall term, the program might have been made to look ineffective because depression often worsens with the oncoming of the winter holiday season, as compared to the early fall.

Seventh, the threat of *selection differences* arises when the composition of the participants measured at pretest differs from the composition of the participants measured at posttest, which could occur if some participants fail to complete the posttest measurement because they leave the study early (i.e., *attrition* or *experimental mortality*). For example, those students whose depression improved the most may have dropped out of the program without completing the posttest. If so, the program would look less effective than it was. Alternatively, students might drop out of the program before completing the posttest because they were frustrated by their lack of progress. If so, the program would look more effective than it was.

The eighth threat to internal validity is *chance*, which means the difference from pretest to posttest is due to random fluctuations, including random measurement error in the pretest and posttest assessments. This threat is what statistical significance tests and confidence intervals address.

Statistical Analysis of Data from the One-Group Pretest–Posttest Design

The statistical analysis of data from the one-group pretest–posttest design consists of a simple paired-sample *t*-test (with a confidence interval as a recommended accompaniment), comparing the difference between the pretest and posttest scores averaged across participants. A simple comparison of the mean of the pretest scores to the mean of the posttest scores using whatever data are available would be susceptible to the effects of selection differences. Using paired-sample differences avoids the effects of selection differences because the same participants are being compared at both pretest and posttest. However, a paired-sample *t*-test does not address any of the other threats to internal validity besides the effects of chance. In addition, if some participants drop out of the study, the results of a paired-sample pretest–posttest comparison might not be generalizable to the population of all the participants who began the program.

Design Embellishments

A simple one-group pretest–posttest design can be embellished by adding a non-equivalent dependent variable, which is a variable collected on the same participants at the time when the original pretests and posttests were collected. A non-equivalent dependent variable is not expected to be influenced by the treatment but is expected to be influenced by the effects of one or more of the same threats to internal validity of the original pretest and posttest comparison. For example, if a treatment is meant to influence verbal ability but not mathematical ability, while history effects are expected to influence both measures, mathematical ability could serve as a non-equivalent dependent variable. A relatively small difference between pretest and posttest measures on the non-equivalent dependent variable (e.g., mathematical ability) suggests that the relevant threats to internal validity (e.g., history effects) are not substantially present to bias the analysis of the data from the original pretest and posttest measures (e.g., verbal ability).

An alternative (or additional) embellishment to a one-group pretest–posttest design is to add a pre-pretest before the pretest, where no treatment intervenes between the pre-pretest and the pretest. A small difference between the pre-pretest and the pretest can suggest that certain threats to interval validity (such as maturation, testing, and cyclical changes) are not operating substantially, hence improving the credibility that the difference from the pretest to posttest is due to the intervening treatment rather than to threats to internal validity.

Further, pre-treatment measures and additional post-treatment measures could be added over time. Doing so would create an interrupted time-series design (discussed in a subsequent section of this chapter). A comparison group of participants could also be added to the simple one-group pretest–posttest design wherein the participants in the comparison group are assessed on both the pretest and posttest but do not receive the treatment. Such a design becomes a non-equivalent group design (see a subsequent section of this chapter as well as Chapter 15 of this volume).

Recommendations

A one-group pretest–posttest design can be quick and easy to use. The problem is that the design can suffer from a range of threats to internal validity, as described herein. In many cases, threats to internal validity will be sufficiently plausible that the design will not provide credible estimates of treatment effects. In such cases, it would be better to consider alternative research designs. However, conditions can arise where the previously listed threats to internal validity are implausible, so estimates of treatment effects will be credible. For example, short time intervals between pretest and posttest can allow little time for either history effects or maturation to threaten the interpretation of results. Objective measuring instruments can reduce the likelihood of instrumentation effects. Environments where assessments using similar measuring instruments are routine can minimize testing effects, and so on. Eckert (2000) provides a convincing example of a short-term instructional intervention where such conditions likely produced highly credible results using a one-group pretest–posttest design. The point is that this type of design can be useful under the right circumstances, but researchers should be ever vigilant for the presence of plausible threats to internal validity.

Non-equivalent Group Designs

In the prototypical non-equivalent group design, two groups of participants are assessed on both pretest and posttest measures, where one group (the treatment condition) receives the treatment, and the other group (the comparison condition) receives either no treatment or an alternative treatment. Non-random assignment of participants to treatment conditions (which makes the design a quasi-experiment) might mean participants self-select into their desired treatment. Or someone other than the participants (e.g., a program administrator or researcher) might assign participants to treatment conditions. The effect of the treatment is estimated by comparing the performances in the two treatment groups. In most cases, the pretest is chosen to be operationally identical to the posttest. For example, two tests of reading skills are operationally identical if they assess the same underlying abilities using parallel instruments. It can also be advantageous in many cases to collect a variety of additional pretest measures.

Examples of the Non-equivalent Group Design

Non-equivalent group designs have been widely used to assess the effects of educational interventions, such as programs to foster the development of intellectual and/or social skills in early childhood (Goplan at el., 2020). In another educational example, Aiken et al. (1998) assessed the effects of a remedial writing program for first-year students in college. Paluck and Green (2009) surveyed the use of non-equivalent group designs in the psychological literature for assessing interventions for reducing prejudice. Heinsman and Shadish (1996) compared the results from non-equivalent group designs to the results from randomized experiments in assessing the effects of coaching on standardized testing, grouping students in classrooms according to their abilities, educating medical patients prior to surgery, and preventing adolescents from engaging in drug abuse. Lehman et al. (1988) used a non-equivalent group design to assess the effects of training in different academic disciplines on reasoning about statistical and methodology principles. Eysenck (1952) used the results of non-equivalent group designs to argue that the results of one-group pretest–posttest designs were not reliable when used to assess the effects of psychotherapy.

Threats to Internal Validity

Selection differences are initial differences between the participants in the two treatment conditions. As they can bias the estimate of a treatment effect, selection differences are a primary threat to the internal validity of non-equivalent group designs. For example, if participants in one treatment condition were more capable at the start than the participants in the other treatment condition, differences on the posttest might be due to such initial differences rather than to the effects of the different treatments. Selection differences are always present in non-equivalent group designs; their effects on the posttest measures must always be addressed when estimating treatment effects. The pretest measures in the two treatment conditions are used to take account of the effects of selection differences on the posttest measures.

Other threats to internal validity, besides selection differences, can also arise. For example, external events could intervene differentially across the groups (labeled *differential history effects*) and bias the difference between the groups on the posttest. For example, consider a non-equivalent group design used to assess the effects of an in-school reading program. A differential history effect would be present if the parents of the children in the in-school reading program also enrolled their children in an out-of-school reading program, while the parents of the children in the comparison condition did not.

The means of addressing differential history effects are the same in quasi-experiments as in randomized experiments. In contrast, non-random selection differences are always present in non-equivalent group designs and the means of coping with them are different in non-equivalent group designs than in randomized experiments. As a result, the focus in the methodological literature on non-equivalent group designs (as in the present chapter) is on addressing the potentially biasing influence of selection differences.

Statistical Analysis of Data from the Non-equivalent Group Design

Various methods have been proposed for analyzing data from non-equivalent group designs with the purpose of taking account of the potentially biasing effects of selection differences. The simplest method is to estimate treatment effects by comparing the treatment groups using the average differences from pretest to posttest. A treatment effect is estimated to be present if one group changed more than the other group, from pretest to posttest. Such an analysis, which assumes the average change over time under the two treatment conditions would remain the same in the absence of a treatment effect, is called a change-score analysis or a differences-in-differences analysis (Angrist & Pischke, 2015).

The obvious weakness of such an analysis is that the treatment effect estimate will be biased if the two treatment groups change from pretest to posttest at a different rate, in the absence of a treatment effect. For example, if the two treatment groups start out responding differently on the pretest, that initial difference might grow over time (even in the absence of a treatment effect) because it is often the case that wealth begets wealth ("the rich get richer").

An alternative to a change-score or differences-in-differences analysis is an analysis based on matching participants from the two treatment groups on their pretest scores. A treatment effect would then be estimated as a difference on the posttest measures between participants from the two groups who were matched on their pretest scores. Matching can be accomplished by either physical or statistical procedures. In physical matching, participants in the two treatment groups are either paired up or put into blocks based on their pretest scores. For example, a participant from the treatment condition who had a pretest score of, say, 75 could be matched with a participant from the comparison condition who also had a pretest score of 75 or close to 75. Alternatively, participants from the treatment group could be matched in blocks where each block contains multiple participants with similar pretest scores. The effect of the treatment is then assessed by calculating the average differences in posttest scores for participants matched or blocked on the pretest scores. Such a procedure addresses the effects of selection differences between participants in the treatment conditions on the measured pretest scores.

In contrast to physical matching, statistical matching can be accomplished using analysis of covariance (ANCOVA), which is a special case of multiple regression. In ANCOVA, the scores on the posttest are regressed onto both the pretest scores and a variable representing treatment assignment – an indicator variable where the value of 1 denotes membership in the treatment condition and 0 otherwise. The regression coefficient for the variable representing treatment assignment is the estimate of the treatment effect. The ANCOVA statistically manipulates the data to assess the treatment effect by comparing the posttest scores of participants from the treatment groups who are matched on the pretest. The difference between a physical matching/ blocking procedure and the ANCOVA is that the matching in ANCOVA is done mathematically rather than by physically putting participants into pairs or blocks; otherwise, the logic underlying the two strategies is the same.

Both the physical and statistical matching analyses can be implemented using multiple pretest measures so as to eliminate the effects of selection differences on all of them. But both analyses become progressively more complicated as the number of pretest measures on which participants are to be matched increases. An alternative to incorporating all the pretest measures individually is to use a single score called the estimated propensity score that predicts assignment to a treatment condition based on the pretest variables (Rubin, 2005). Estimated propensity scores are then used in either the physical matching/blocking analysis or the ANCOVA. To the extent that the true propensity scores have been well estimated, an analysis using the estimated propensity scores simultaneously takes account of the effects of selection differences on all the measured pretest scores that were used in creating the propensity scores.

The weaknesses of both physical matching/blocking procedures and the ANCOVA are twofold; all necessary pretest measures might not be available, and biases can be introduced by measurement error in the pretest measures. If selection differences exist between the participants in the two treatment groups on pretest variables that have not been included in the analysis or if the pretest variables included in the analysis have been measured with error, both a physical and statistical matching analysis can be biased because effects of selection differences remain. Under many, if not most, circumstances, it is difficult to be convinced that the effects of all selection differences have been adequately removed in the analyses.

Design Embellishments

The non-equivalent group design can be usefully elaborated in at least two ways. First, a pre-pretest can be collected in both treatment groups. The pre-pretest is assessed before the original pretest in the non-equivalent group design. The data from the pre-pretest and the original pretest are analyzed as if they were data from a non-equivalent group design (to produce what is called a dry-run analysis), where null results are expected because the treatment has not yet been introduced. Null results from the dry-run analysis suggest that the effects of selection differences have been adequately taken into account and increase the credibility that the effects of selection differences have been adequately taken into account in the original non-equivalent group design, using the same statistical analysis.

The second elaboration is to add a non-equivalent dependent variable, where such a variable is chosen to be free of the effects of the treatment but to share the same effects as the original outcome measure of selection differences. Again, finding null results for the non-equivalent dependent variable increases the researcher's confidence that the same statistical analysis removes the effects of selection differences using the pretest and posttest from the original non-equivalent group design.

Recommendations

The non-equivalent group design is a one-group pretest–posttest design with an added comparison condition, which helps render implausible some threats to internal validity that are often plausible in one-group pretest–posttest designs. However, initial selection differences remain a serious threat to the internal validity of the non-equivalent group design. Unfortunately, no statistical procedure can be guaranteed to fully take account of the effects of initial selection differences. That is, modeling the effects of selection differences (so their effects can be adequately taken into account) is fraught with potential error and uncertainty. Therefore, researchers can be left with substantial doubt about the likely size of treatment effects when using a non-equivalent group design.

Assigning participants to treatment conditions at random produces a randomized experiment rather than a non-equivalent group design; this is one of the recommended ways to cope with the potentially biasing effects of initial selection differences. Estimates of treatment effects from randomized experiments can still be biased by problems such as *differential attrition* and *non-compliance* to treatment assignment. Differential attrition means different types of participants from the treatment conditions leave the study before posttest measures are collected. Non-compliance means some participants do not receive their assigned treatment conditions (Sagarin et al., 2014). Careful implementation of randomized experiments, however, can help to minimize such potential sources of bias. In addition, estimates of treatment effects from randomized experiments tend to be more precise than estimates of treatment effects from non-equivalent group designs (because random assignment makes treatment assignment uncorrelated with pretest variables, thereby avoiding multicollinearity). The bottom line is that the results of randomized experiments tend to be more trustworthy than the results of non-equivalent group designs and so a randomized experiment should be considered as an alternative to a non-equivalent group when feasible.

When non-equivalent group designs are used, several steps should be taken to obtain the most credible results: make the treatment and comparison groups as similar as possible initially, ascertain the nature of inevitable selection differences, assemble a wide range of pretest measures to assess selection differences and employ the measures in a variety of credible statistical procedures to adjust for the effects of selection differences, and add design elaborations, such as non-equivalent dependent variables (Cook et al., 2009; Reichardt, 2019). With careful implementation under such conditions, non-equivalent group designs are capable of producing results similar to the results from randomized experiments (Cook et al., 2008). Hence, the results from non-equivalent group designs can be credible, but that will not likely be accomplished without careful, rather than casual, implementation of the research design.

Regression Discontinuity Designs

In a regression discontinuity design, participants are assessed on a quantitative assignment variable (QAV) before the treatment conditions are introduced, and a cut-off value on the QAV is used to assign participants to the treatment conditions they are to receive. If participants have QAV scores above the cut-off value, they are assigned to one of the treatment conditions. If participants have QAV scores below the cut-off value, they are assigned to the other treatment condition.

If the treatment is meant to address a problem from which participants might suffer, the QAV could be a measure of need for the treatment, with those most needy being assigned to the treatment. For example, if the treatment were meant to address a learning deficit, the QAV might be a measure of academic performance, and the treatment would be given to those who score lowest on the QAV. Alternatively, the treatment could be a pay-off (e.g., a college scholarship) for outstanding prior performance, with those scoring highest on the QAV measure of performance receiving the treatment. Other QAV measures have also been used, such as when treatment is awarded based on when study participants file an application to receive services or when the treatment is given to those who have reached a certain age when the treatment begins. After their assignment to treatment conditions based on the QAV, the participants receive the different treatments, and the outcomes are subsequently assessed.

Figure 14.1 plots the scores on the outcome measure versus the QAV scores for hypothetical data from a regression discontinuity design. The treatment condition contains those participants with QAV scores below the cut-off value of 60, as indicated by the dashed vertical line in the figure. In contrast, the comparison condition contains those participants with QAV scores above the cut-off value. For the regression discontinuity analysis, the scores on the outcome measure are regressed onto the QAV scores in each treatment condition. These regression lines are indicated by the solid lines in the figure, where the dotted lines are extensions of the regression lines showing what the regression lines would be if extrapolated from one side of the cut-off value to the other.

To estimate the effect of the treatment, the regression lines in the two treatment groups are compared. Note in Figure 14.1 that there is a discontinuity between the regression lines in the two groups at the cut-off value. Such a discontinuity is taken as evidence of a treatment effect. In Figure 14.1, there is a positive treatment effect because the regression line in the treatment condition is raised above the regression line in the comparison condition. The treatment effect would be estimated to be negative if the regression line in the treatment condition had been lower than the regression line in the comparison condition. In the absence of a treatment effect, there would be no discontinuity in the regression lines at the cut-off value. That is, without a treatment effect, the regression lines would fall on top of each other. The logic of the design is that a treatment effect will produce a discontinuity right at the cut-off value because the treatment assignment changes right at the cut-off value.

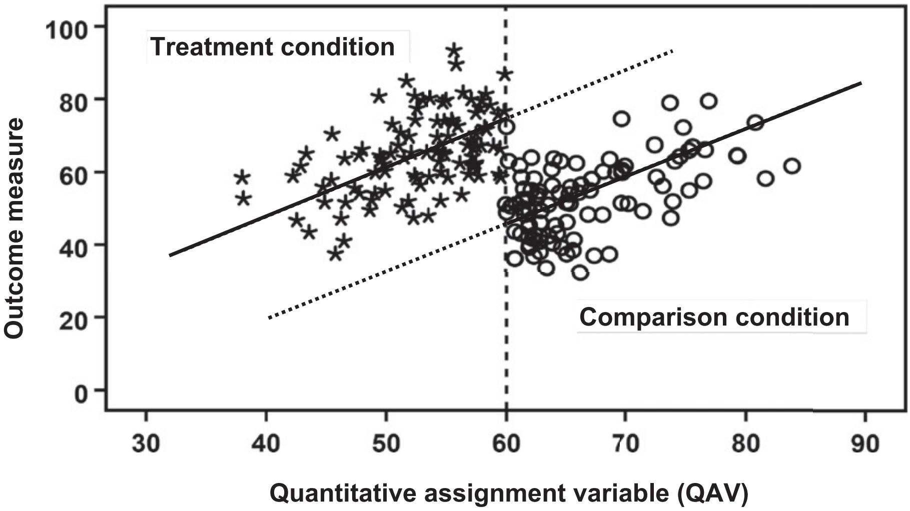


Figure 14.1 *Hypothetical data showing a positive treatment effect in a regression discontinuity design.*

In Figure 14.1, the effect of the treatment is constant across the QAV scores because the regression lines are parallel – the treatment raises the outcome scores equally across participants with different QAV scores. It is possible that the regression lines are not parallel. In that case, the effect of the treatment interacts with the QAV scores. If the slope of the regression line in the treatment condition is steeper than the slope of the regression line in the comparison condition, the treatment effect is estimated to be greater for participants with higher QAV scores than for participants with lower QAV scores, and vice versa. A change in level is said to arise when the treatment effect produces a discontinuity in the regression lines at the cut-off value. A change in slope is said to arise when the treatment effect produces non-parallel regression lines that represent an interaction effect of the treatment with the QAV scores.

Examples of the Regression Discontinuity Design

The regression discontinuity design was originally devised by Thistlewaite and Campbell (1960), who used it to assess the effects of a national designation of academic merit on students' subsequent performances and career choices. Since then, the design has been used sporadically over time and across disciplines (Cook, 2008). Since the 1990s, the regression discontinuity design has been widely used in research in economics, as demonstrated by the numerous citations in Lee & Lemieux (2010). Trochim (1984) reported on the frequent use of the regression discontinuity design to assess the effects of compensatory educational programs on student achievement from the war on poverty programs that began in the 1960s. And as seen in Henry et al. (2010) and Henry and Harbatkin (2020), the design continues to be used in the evaluation of the effects of educational programs. Both Braden and Bryant (1990) and Matthews et al. (2012) discussed the use of the regression discontinuity design in assessing the effects of educational programs directed to gifted and talented children. Berk et al. (2010) found that the results from a regression discontinuity design produced comparable results to a randomized experiment in assessing the effects on recidivism of the relaxation of supervision following participants' release from prison. Mark and Mellor (1991) used a regression discontinuity design to assess the effects of negative life events on the psychological variable of hindsight bias.

Threats to Internal Validity

As noted, the analysis of data from a regression discontinuity design entails fitting regression lines between the outcome scores and the QAV measure and looking for a discontinuity in the regression lines in the two treatment groups at the cut-off value and/or a change in the slope of the two regression lines. Bias can arise when the regression surfaces are estimated to be straight lines, but the true regression surfaces are curvilinear. Such misfits of the regression surface introduce the threat to internal validity due to *curvilinearity.* In the presence of curvilinearity, a straight-line fit can produce an apparent change in the level or slope (or both). To avoid such biases, the regression surfaces being estimated must fit the true curvilinear shape, as described below.

Non-compliance to treatment conditions can also produce biases in the estimate of a treatment effect in a regression discontinuity design. Non-compliance arises when participants assigned to one treatment condition receive the alternative treatment condition instead. Some participants assigned to the treatment condition might not show up to receive the treatment and/or some participants originally assigned to the comparison condition might finagle their way into receiving the treatment, either within the confines of the study or by seeking a similar treatment outside the implementation of the study. Both forms of non-compliance produce a fuzzy regression discontinuity design. Non-compliance to treatment conditions can bias the estimates of treatment effects to the extent that those who don't comply with treatment assignment differ in their outcomes compared to those who adhere to their treatment assignments. Means of adjusting for such potential biases entail estimating the treatment effect for those participants who comply with their treatments as assigned by the QAV cut-off value (in essence, ignoring those who don't comply). But these methods require strict assumptions about the nature of non-compliance (Imbens & Lemieux, 2008).

Differential attrition can occur in a regression discontinuity design when different types of people drop out of the two treatment conditions. Differential attrition can bias the estimates of treatment effects to the extent that the types of participants who drop out of the treatment condition tend to differ from the types of participants who drop out of the comparison condition. Researchers address the effects of differential attrition in the regression discontinuity design using the same methods as in a randomized experiment.

Manipulation of the QAV scores might allow participants to gain access to a different treatment condition than the one to which they should have been assigned. A participant with a true QAV score lying above the cut-off value might want to be in the treatment condition for participants with QAV scores below the cut-off value and therefore misrepresent (or have someone else misrepresent) their QAV score, or vice versa. Manipulation of the QAV scores can bias the treatment effect estimates because certain types of participants can tend to be differentially enrolled in the different treatments. Discontinuities at the cut-off value in the frequency distribution of the QAV scores suggest either differential attrition or manipulation of the QAV scores (McCrary, 2008).

Statistical Analysis of Data from the Regression Discontinuity Design

Two approaches for fitting the regression lines in the regression discontinuity design are most common: global and local regression analysis. The difference between the two analyses lies in the amount of data used to fit the regression lines. Global regression uses the data from all participants. Local regression uses data only from participants whose QAV scores are closest to the cut-off value.

Global analysis can be performed using ANCOVA (see the previous section). In the simplest ANCOVA model, the outcome measure is regressed onto both a rescaled QAV measure and an indicator variable representing treatment assignment. The rescaled QAV measure is created by taking the difference between each participant's

QAV score and the cut-off value. An ANCOVA model with these two specified independent variables would fit the data displayed in Figure 14.1. The regression coefficient for the indicator variable is the estimate of the treatment effect for a change in level.

When the regression lines are not parallel (i.e., when a treatment interaction is present), an interaction variable is added to the ANCOVA. The interaction variable is formed by multiplying the indicator variable and the rescaled QAV measure. The interaction variable is then added as another independent variable in the ANCOVA model. The regression coefficient for the interaction variable is the estimate of the treatment effect due to a change in slope; the regression coefficient for the indicator variable is the estimate of the treatment effect (i.e., the discontinuity in level) assessed at the cut-off value on the QAV.

The most common way to fit a curvilinear regression surface (and thereby to try to avoid bias that could arise due to the presence of curvilinearity) is to add polynomial terms to the ANCOVA model. For example, to fit a quadratic curvilinear shape, the rescaled QAV measure is squared and added to the ANCOVA analysis. Higher-order polynomial terms can also be added to the ANCOVA model in a similar fashion to take account of more complex curvilinear shapes. Unfortunately, there is no guarantee a polynomial model that can be reasonably crafted will fit the curves that exist in the data very well. Various steps, such as checking the pattern of residuals from model fits, should be taken to diagnose misfits of the regression surface.

In local regression analysis, a distance (called a bandwidth) is chosen, and those participants with QAV scores further from the cut-off value than the bandwidth are excluded from the analysis (Jacob et al., 2012). The ANCOVA models specified above are then fit to the included data. The logic for excluding data in local regression analysis is based on a trade-off between the power of the statistical analysis and bias. Global regression analysis is more powerful than local regression analysis because global analysis uses all the data, but local analysis is likely to be less biased than global analysis in the presence of curvilinearity. The local approach is likely to be less biased by curvilinearity because the regression surface is likely to be less curved over a shorter, than longer, stretch of QAV scores. That is, as the range of QAV scores gets narrower and narrower, the regression surface tends to become more and more linear even in the presence of curvilinearity. Hence, there is less likelihood of bias from misfitting curvilinearity in local, as compared to global, analysis.

Design Embellishments

The basic regression discontinuity design can be usefully elaborated in at least two ways. First, a pre-treatment measure that is operationally identical to the outcome measure can be added to the design. Second, the same outcome and QAV measures can be collected from a group of participants where the treatment is not made available (called a non-equivalent comparison group). Then, the data from a regression discontinuity design can be analyzed simultaneously with either an operationally identical pretest or a non-equivalent comparison group to increase the credibility and power of the results. The logic of such analyses is that the operationally identical pretest and the non-equivalent comparison group add extra data as well as being a means of assessing assumptions of the analyses that can bolster the results from the original regression discontinuity design (Reichardt, 2019).

Recommendations

The fact that the regression discontinuity design requires participants be assigned to treatment conditions based on their scores on a quantitative variable reduces the circumstances in which the design can be implemented compared, for example, to the non-equivalent group design. Both designs suffer from the presence of selection differences between the participants in the different treatment conditions, but the QAV in the regression discontinuity design can provide a plausible means of adjusting for the effects of selection differences. In contrast, the pretest measures in the non-equivalent group design may or may not adequately represent all important selection differences. As a result, the regression discontinuity design is usually thought to provide more credible estimates of treatment effects than the non-equivalent group design. The potential benefits of the regression discontinuity design, especially as compared to the non-equivalent group design, suggest that researchers be attuned to opportunities in which the design can be implemented. In addition, recipients of treatments are sometimes assigned to treatment conditions according to a quantitative variable, even when that assignment is not initiated by the researchers. Researchers should be on the lookout for such desirable circumstances so a regression discontinuity analysis can be performed, and the benefits of the design obtained.

Randomized experiments are still considered by many to be preferable to regression discontinuity designs because randomized experiments require fewer assumptions in data analysis. Unbiased estimation of treatment effects in a randomized experiment does not require that a regression surface be properly fit to potentially curvilinear data, while the analysis of data from the regression discontinuity design imposes that requirement. In addition, even under ideal conditions, regression discontinuity designs can require more than twice as many participants to have the same statistical power as randomized experiments (Goldberger, 2008). But regression discontinuity designs can often be an adequate replacement for randomized experiments when randomized experiments can’t be well implemented (Cook et al., 2008). The regression discontinuity design appears to be less well known than other designs, and researchers are advised to become more familiar with the design so it can be implemented and its benefits reaped. Researchers should recognize that, with careful implementation, the regression discontinuity design can produce highly credible estimates of treatment effects.

Interrupted Time-Series Designs

In the prototypical interrupted time-series design, observations are collected at multiple points spaced out in time before a treatment is implemented as well as at multiple points spaced out in time after a treatment is implemented. Using regression analysis, the researcher models the trends in the observations over time separately before and after the treatment is implemented. The trend in the data before the treatment is implemented is continued forward in time into the region of the data following the treatment. This projected trend in performance is then compared to the actual trend in performance following the introduction of the treatment. The treatment effect is assessed based on differences between the two trends.

In Figure 14.2, scores on an outcome measure are plotted across time for hypothetical data in an interrupted time-series design. The dots in the figure are the data points. The trends in the data, as fit by regression analysis for both the pre-treatment and post-treatment observations, are indicated by the solid diagonal lines. The regression line from the pre-treatment data is extrapolated forward in time to produce the dotted regression line in the figure.

The data in Figure 14.2 reveal a positive treatment effect because the trend following the introduction of the treatment that is estimated based on the pre-treatment data falls below the trend that arises in the post-treatment data. As a result of the treatment effect, there is an interruption in the trends in these two regression lines when the treatment is introduced (as indicated by the vertical dashed line) – thus the name interrupted time-series design. In contrast, a treatment would be estimated to have no effect if the projected and actual post-treatment trends fell on top of one another (i.e., there is no interruption in the regression lines at the time the treatment is introduced or any differences between the two trends after the treatment was introduced).

Because the projected and actual post-treatment trends are parallel in Figure 14.2, the treatment effect is estimated to be constant over time. It is also possible that the projected and actual post-treatment trends are not parallel, in which case the treatment effect is estimated to vary over time (either increasing or

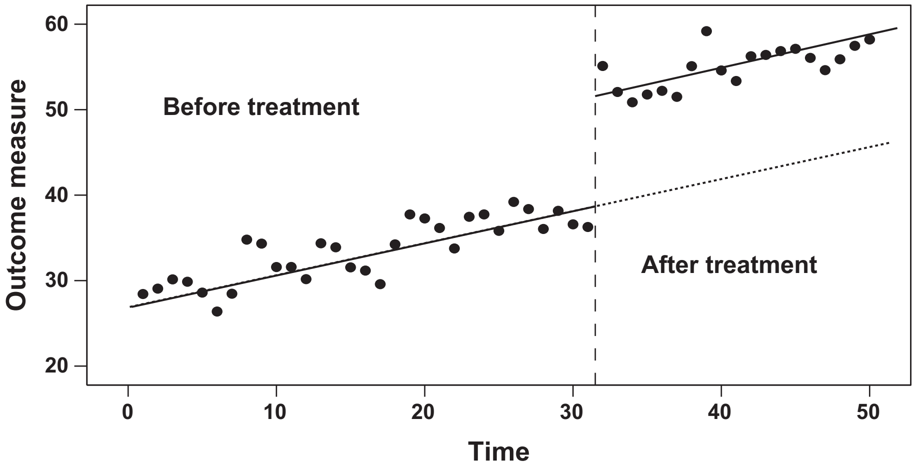


Figure 14.2 *Hypothetical data showing a positive treatment effect in an interrupted time-series design.*

decreasing). For example, a change in diet might have an increasingly positive effect on health over time. Alternative patterns of treatment effects are also possible. Among many other possibilities, the treatment effect could be delayed wherein its effect starts sometime after the treatment is first introduced. Or a treatment effect could first grow and then diminish over time. The interrupted time-series design can be implemented using either a single participant or multiple participants. The data in Figure 14.2 are from a single participant. If there were multiple participants, there would be multiple observations at each time point.

Examples of the Interrupted Time-Series Design

The famous Hawthorne experiments of the effects of working conditions on worker productivity used interrupted time-series designs (McCleary & McDowall, 2012). The interrupted time-series design is widely used in applied behavior analysis, often under the name single-case designs, wherein the effects of interventions on problem behaviors are assessed (Kazden, 2011; Nugent, 2010). Bloom (2003) used interrupted time-series designs to assess the effects of whole-school reforms on student performance. The interrupted time-series design has also been used to assess the effects of diversion programs in the justice system on juvenile crime recidivism (Lipsey et al., 1981) and the effects of an incentive program on lottery sales (Reynolds & West, 1987). Hudson et al. (2019) found 116 studies in the field of health care, published in 2015 alone, that used the interrupted time-series design. Palmgreen (2009) reports on the use of an interrupted time-series design to assess the effects of televised public service announcements on attitudes about illicit drug use in sensation-seeking adolescents.

Threats to Internal Validity

A primary source of bias in the analysis of interrupted time-series data is the misspecification of the trends in the data over time. For example, both the pre-treatment and post-treatment trends (both actual and projected) are linear in Figure 14.2. It is possible, instead, that some or all the trends in the data are curvilinear in the absence of a treatment effect (e.g., exponential growth in infections due to a pandemic) and the modeled trends should also be curvilinear, if bias is to be avoided. Unfortunately, accurately modeling curvilinearity is not guaranteed by any statistical procedure but must rest, instead, on assumptions of the analysis.

Bias in the analysis of data from an interrupted time-series design can also arise when other changes, besides the treatment, occur when the treatment is introduced. For example, either history or instrumentation effects could arise and introduce spurious interruptions in the time-series data. A history effect is present in an interrupted time-series design when an influential event other than the treatment arises at the same time the treatment is implemented. An instrumentation effect would arise if a change or recalibration in the measuring instrument was introduced at the same time the treatment was implemented.

Statistical Analysis of Data from the Interrupted Time-Series Design

In an interrupted time-series design, both the observed and projected trends for the post-treatment data are modeled using the same ANCOVA procedure as presented in the analysis of data for the non-equivalent group and regression discontinuity designs, with the following exception (Reichardt, 2019; Somers et al., 2013). The statistical analysis of interrupted time-series data is complicated by the likely presence of correlations between observations across time. For example, a person with above-average performance in school one semester is more likely than not to have above-average performance in school the next semester, as compared to other students. Such a correlation between school performances over time reflects a positive autocorrelation. Negative autocorrelations are also possible. For example, a sleepless night might tend to be followed by more restful sleep the next night. The problem for the interrupted time-series design is that time-series data tend to be autocorrelated while standard statistical analyses (such as ANCOVA) assume data are free from autocorrelations. Performing standard analyses in the presence of autocorrelation tends to bias the analyses so that statistical power and precision are misrepresented. The alternative is to add specialized statistical procedures, such as auto-regressive, moving average (ARMA) models, to the ANCOVA analysis. Based on patterns in the data, ARMA models specify a structure for the nature of the autocorrelation and then adjust for the biasing effects of autocorrelation, assuming the structure has been properly specified.

Design Embellishments

The prototypical interrupted time-series design (as depicted in Figure 14.2) can be elaborated by adding data from a comparison time series of observations. Adding comparison data is highly recommended and can take one of two forms. First, a time series of comparison data can be derived from a non-equivalent-dependent variable collected on the same participants as in the original interrupted time-series design. Second, time-series data can be added from a comparison group of participants that does not receive the treatment. In either case, the comparison data are selected so they exhibit no treatment effect but are influenced by the same threats to validity as in the original interrupted time-series data. The credibility of the interrupted time-series analysis is strengthened to the extent to which the comparison data evidence no interruption in the trend of the observations after the treatment is introduced. Conversely, an interruption in the comparison data at the time of the treatment suggests that a threat to internal validity (e.g., history effects) is operating; this could also account for an interruption in the original interrupted time-series data. Thus, finding no interruption in the data from the comparison data suggests the threat to internal validity is also not operating in the original interrupted time-series data.

Recommendations

Because it adds observations over time, the interrupted time-series design extends the simple one-group pretest–posttest design. The added observations can reduce the plausibility of threats to validity. For example, the added observations allow the effects of maturation to be modeled and thereby removed. Similarly, additional pre-treatment observation can make regression toward the mean apparent, if it is present, and multiple pre-treatment observations can reduce the effects of testing. Adding observations over time can also add to the credibility of other quasi-experimental designs as well. As a result, researchers should routinely consider adding observations over time whenever estimating treatment effects.

Archives of data, collected by those besides the researcher, can sometimes provide the observations necessary to implement an interrupted time-series design and should be considered as a potentially valuable resource when estimating treatment effects. The bottom line is that the interrupted time-series design can provide highly credible estimates of treatment effects when well implemented and should become a well-recognized option in researchers' collection of research designs.

Conclusion

The one-group pretest–posttest, the non-equivalent group, the regression discontinuity, and the interrupted time-series designs are the most common quasi-experimental designs (see Reichardt, 2019, for other designs and for additional details and suggestions about the design and analysis of quasi-experiments). In addition, further embellishments to these designs, besides the ones we have described, can also be added to quasi-experiments. Rather than implementing one of the prototypical designs described herein, quasi-experiments should be fashioned, by choosing among the many possible options, to best fit the research circumstances. The types of designs that can be implemented are unlimited, given sufficient researcher ingenuity. We have described the underlying logic of the four prototypical quasi-experiments, examples of each design, the most likely threats to internal validity to these designs, and how to analyze data from these designs to cope with such threats. The results can be generalized to other types of quasi-experiments and design embellishments.

Different research designs have different strengths and weaknesses. In creating a research design, researchers must balance the relative strengths and weaknesses of the various design options. Randomized experiments are often considered the best design, but randomized experiments are not immune to bias due to such things as differential attrition and treatment non-compliance. In addition, randomized experiments can't always be implemented because of either ethical or practical obstacles. Even when randomized experiments can be used, they might have to be implemented with smaller sample sizes than quasi-experiments. As a result, although randomized experiments tend to produce more precise estimates of treatment effects than quasi-experiments when sample sizes are the same, quasi-experiments can sometimes produce more precise estimates than randomized experiments when implemented in practice. In at least some circumstances, quasi-experiments can be the better design option than the presumed gold-standard of randomized experiments.

References

Aiken, L. S., West, S. G., Schwalm, D. E., Carroll, J. L., & Hsiung, S. (1998). Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation: Efficacy of a university-level remedial writing program. *Evaluation Review*, *22*(2), 207–244.

Angrist, J. D. & Pischke, J-S. (2015). *Mastering 'Metrics: The Path from Cause to Effect*. Princeton University Press.

Arum, R. & Roksa, J. (2010). *Academically Adrift: Limited Learning on College Campuses*. University of Chicago Press.

Berk, R., Barnes, G., Ahlman, L., & Kurtz, E. (2010). When second best is good enough: A comparison between a true experiment and a regression discontinuity quasi-experiment. *Journal of Experimental Criminology*, *6*(2), 191–208.

Bloom, H. S. (2003). Using "short" interrupted time-series analysis to measure the impacts of whole-school reforms: With application to a study of accelerated schools. *Evaluation Review*, *27*(1), 3–49.

Braden, J. P. & Bryant, T. J. (1990). Regression discontinuity designs: Applications for school psychologists. *School Psychology Review*, *19*(2), 232–239.

Cook, T. D. (2008). "Waiting for life to arrive": A history of the regression–discontinuity designs in psychology, statistics and economics. *Journal of Econometrics*, *142*(2), 636–654.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, *27*(4), 724–750.

Cook, T. D., Steiner, P. M., & Pohl, S. (2009). Assessing how bias reduction is influenced by covariate choice, unreliability and data analysis mode: An analysis of different kinds of within-study comparisons in different substantive domains. *Multivariate Behavioral Research*, *44*(6), 828–847.

Eckert, W. A. (2000). Situational enhancement of design validity: The case of training evaluation at the World Bank Institute. *American Journal of Evaluation*, *21*(2) 185–193.

Eysenck, H. J. (1952). The effects of psychotherapy: An evaluation. *Journal of Consulting Psychology*, *16*(5), 319–324.

Goldberger, A. S. (2008). Selection bias in evaluation treatment effects: Some formal illustrations. In T. Fomby, R. C. Hill, D. L. Millimet, J. A. Smith, & E. J. Vytlacil (eds.), *Modeling and Evaluating Treatment Effects in Economics* (pp. 1–31). JAI Press.

Goplan, M., Rosinger, K., & Ahn, J. B. (2020). Use of quasi-experimental research designs in education research: Growth, promise, and challenges. *Review of Research in Education*, *44*(1), 218–243.

Heinsman, D. T. & Shadish, W. R. (1996). Assignment methods in experimentation: When do nonrandomized experiments approximate answers from randomized experiments? *Psychological Methods*, *1*(2), 154–169.

Henry, G. T., Fortner, C. K., & Thompson, C. L. (2010). Targeted funding for educationally disadvantaged students: A regression discontinuity estimate of the impact on high school student achievement. *Educational Evaluation and Policy Analysis*, *32*(2), 183–204.

Henry, G. T. & Harbatkin, E. (2020). The next generation of state reforms to improve their lowest performing schools: An evaluation of North Carolina's school transformation intervention. *Journal of Research on Educational Effectiveness*, *13*(4), 702–730.

Hudson, J., Fielding, S., & Ramsay, C.R. (2019). Methodology and reporting characteristics of studies using interrupted time series design in healthcare. *BMC Medical Research Methodology*, *19*(1), 137.

Imbens, G. W. & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice, *Journal of Econometrics*, *142*(2), 615–635.

Jacob, R., Zhu, P., Somers, M-A., & Bloom, H. (2012). *A Practical Guide to Regression Discontinuity*. Manpower Demonstration Research Corporation.

Kazden A. E. (2011). *Single-Case Research Designs: Methods for Clinical and Applied Settings*, 2nd ed. Oxford University Press.

Lee, D. S. & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, *48*(2), 281–355.

Lehman, D. R., Lempert, R. O., & Nisbett, R. E. (1988). The effects of graduate training on reasoning: Formal discipline and thinking about everyday-life events. *American Psychologist*, *43*(6), 431–442.

Lipsey, M.W., Cordray, D.S., & Berger, D.E. (1981). Evaluation of a juvenile diversion program: Using multiple lines of evidence. *Evaluation Review*, *5*(3), 283–306.

Mark, M. M. & Mellor, S. (1991). The effect of self-relevance of an event on hindsight bias: The foreseeability of a layoff. *Journal of Applied Psychology*, *76*(4), 569–577.

Matthews, M. S., Peters, S. J., & Housand, A. M. (2012). Regression discontinuity design in gifted and talented education research. *Gifted Child Quarterly*, *56*(2), 105–112.

McCleary, R. & McDowall, D. (2012). Time-series designs. In H. Cooper, P. M. Camic, D. L. Long, et al. (eds.), *APA Handbook of Research Methods in Psychology, Volume 2. Research Designs: Quantitative, Qualitative, Neuropsychological, and Biological* (pp. 613–627). American Psychological Association.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, *142*(2), 698–714.

Nugent, W. R. (2010). *Analyzing Single System Design Data*. Oxford University Press.

Palmgreen, P. (2009) Interrupted time-series designs for evaluating health communication campaigns. *Communication Methods and Measures*, *3*(1–2), 29–46.

Paluck, E. L. & Green, D. P. (2009). Prejudice reduction: What works? A review and assessment of research practice. *Annual Review of Psychology*, *60*, 339–367.

Reichardt, C. S. (2019). *Quasi-Experimentation: A Guide to Design and Analysis*. Guilford Press.

Reynolds, K. D. & West, S. G. (1987). A multiplist strategy for strengthening nonequivalent control group designs. *Evaluation Review*, *11*(6), 691–714.

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

Sagarin, B. J., West, S. G., Ratnikov, A., Homan, W. K., & Ritchie, T. D. (2014). Treatment noncompliance in randomized experiments: Statistical approaches and design issues. *Psychological Methods*, *19*(3), 317–333.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton-Mifflin.

Somers, M.-A., Zhu, P., Jacob, R., & Bloom, H. (2013). *The Validity and Precision of the Comparative Interrupted Time Series Design and the Difference-in-Difference Design in Educational Evaluation*. Manpower Demonstration Research Corporation.

St. Pierre, R. G., Ricciuti, A., & Creps, C. (1999). *Synthesis of Local and State Even Start Evaluations*. Abt Associates.

Thistlewaite, D. L. & Campbell, D. T. (1960). Regression–discontinuity analysis: An alternative to the ex-post-facto experiment. *Journal of Educational Psychology*, *51*(2), 309–317.

Trochim, W. M. K. (1984). *Research Designs for Program Evaluation: The Regression–Discontinuity Approach*. SAGE Publications.

https://doi.org/10.1017/9781009010054.015 Published online by Cambridge University Press

15 Non-equivalent Control Group Pretest–Posttest Design in Social and Behavioral Research

Margaret Denny, Suzanne Denieffe, and Kathleen O’Sullivan

Abstract

Experimental research designs feature two essential ingredients: manipulation of an independent variable and random assignment of subjects. However, in a quasi-experimental design, subjects are assigned to groups based on non-random criteria. This design allows for manipulation of the independent variable with the aim of examining causality between an intervention and an outcome. In social and behavioral research, this design is useful when it may not be logistically or ethically feasible to use a randomized control design – the “gold standard.” Although not as strong as an experiment, non-equivalent control group pretest–posttest designs are usually higher in internal validity than correlation designs. Overcoming possible threats to internal and external validity in a non-equivalent control group pretest–posttest design, such as cofounding variables, are discussed in relation to sample selection, power, effect size, and specific methods of data analyses.

Keywords*:* Quasi-experimental Design; Non-equivalent Group; Power; Effect Size; Data Analysis

Introduction

This chapter explores a type of quasi-experimental research design, referred to as a non-equivalent control group pretest–posttest (NECGPP) design, which can be used in social and behavioral sciences. This chapter is an exposition of the steps of the research process required for a NECGPP design. The chapter specifically focuses on the rationale for using a NECGPP design, sample selection, power, effect size, strengths, weaknesses, and the specific methods of data analyses commonly used to analyze data resulting from a NECGPP.

Research Methodology

Research Design

At the core of the research process is the researcher's plan for how the study should unfold. In social and behavioral research, scientists have credited experimental design approaches with great powers of explanation and prediction. For instance, Riley (1967, p. 612) suggests that experimental designs are:

> A powerful [way of] testing hypotheses of casual relationships among variables. Ideally, in the experimental design the investigator throws into sharp relief the explanatory variables in which he is interested, controlling or manipulating the independent variable . . . observing its effect on the dependent variable . . . and minimising the effects of the extraneous variable, which confirmed his results.

The key feature of randomized control designs is the random assignment of participants to groups. As a result, these are viewed as the gold standard in research because the methods allow one to prevent many biases that can be due to demand characteristics/artifacts, the placebo effect, or indeed many other cofounding variables (see Chapter 14 in this volume). However, double-blind studies are not always possible outside laboratory settings (Cook & Campbell, 1979) due to ethical or practical reasons. Experimental science researchers have therefore developed quasi-experimental designs (Benjamin, 1988; see also Chapter 14 in this volume). Quasi means "resembling" or "having some of the features of." As such, a quasi-experiment resembles an experiment; it has some but not all the features of an experiment. Campbell and Stanley (1966, p. 34) initiated the term quasi-experimental to refer to research designs "that lack full control over the scheduling of experimental stimuli, that is, the when and to whom of exposure and the ability to randomize exposures." Campbell and Stanley (1966, p. 2) insist that such designs are ". . . the only way of establishing a cumulative tradition in which improvements can be introduced without the danger of a faddish discard of old wisdom in favor of inferior novelties."

There are two main types of quasi-experimental design: *non-equivalent group designs* and *cohort designs* (Heppner et al., 1992, p. 152). Quasi-experimental designs involve the manipulation of an independent variable (e.g., the introduction of a treatment or intervention; Polit, 2005; Polit & Beck, 2004; Polit et al., 2006). The argument for and against randomized control designs versus quasi-experimental designs still tends to dominate the debate in the approaches to research. Researchers who utilize quasi-experimental designs argue that there is generally little loss of status or cachet when such designs are employed (Daws et al., 2005).

Quasi-experimental Design

The influences that affect choice of design mainly center on cost and contextual issues, such as the setting in which the research is to take place. This includes research where intact samples are used – particularly in education research (e.g., students doing a specific module of study). In quasi-experimental designs, matching characteristics (similar demographics) are often used instead of randomization (Campbell and Stanley, 1966). Quasi-experimental designs are not only concerned with association, which implies covariation, but also with the many different and interlocking relationships between variables and many cofounding variables (Murray, 2003). In the same way, it is not only about finding one factual trend, it can expose several trends (Crotty, 2006).

When randomization is not possible or realistic, quasi-experimental designs are viewed as suitable research design alternatives. Rosenthal and Rosnow (2008) posit that no single research method can totally embrace the complexities of human nature and the extraneous variables that impact most research designs. They call this doctrine *methodological pluralism,* which has its roots not in philosophical contextualism and theoretical ecumenism, but in different methodological operations (Rosenthal & Rosnow, 2008). While quasi-experimental designs are weaker in terms of design, they have their merits because they present uncomplicated findings without the complex restraints of randomized control designs (Hunsley & Lee, 2006), which are not always feasible in social and behavioral research. There are five types of quasi-experimental designs. These are the posttest only design with non-equivalent groups, the pretest–posttest design with non-equivalent groups (i.e., NECGPP), the interrupted time-series design with non-equivalent groups, the pretest–posttest design with switching replication, and the switching replication with treatment removal design (see Chapter 14 in this volume). The focus of this chapter in on the NECGPP design.

Non-equivalent Pretest–Posttest Control Group Design

A NECGPP is a robust design, as it involves selecting a control group of participants who are comparable to the treatment group (i.e., intact groups). For example, in education, we might pick two comparable classrooms or schools that would be intact groups. Alternatively, the researcher can form these intact groups from within their sample. The researcher, for example, is interested in testing the effectiveness of a teaching strategy intervention on learning and students' end of year results. A NECGPP design can statistically control for differences between groups, prior to the commencement of the research by matching characteristics for the treatment and control participants and, at the data analyses phase, by using specific statistical data analytic approaches (Rubin, 1979; Dickinson et al., 1987).

In a NECGPP design, two groups are used: one receives treatment while the other group acts as the control (Figure 15.1). It is important to clarify that there is not a random assignment to these two groups. These groups are intact groups – two existing groups that are already formed are used (e.g., two comparable classrooms or schools). This then maximizes the effectiveness of the design by selecting groups that are as similar as possible so that they will experience and respond similarly to extraneous influences or confounding variables.

According to McMillan (2000), this research design is best suited when participants are in existing groups (intact groups), as is common in educational research (Heppner et al., 1992, 2004; Frank & Gilovich, 1988).

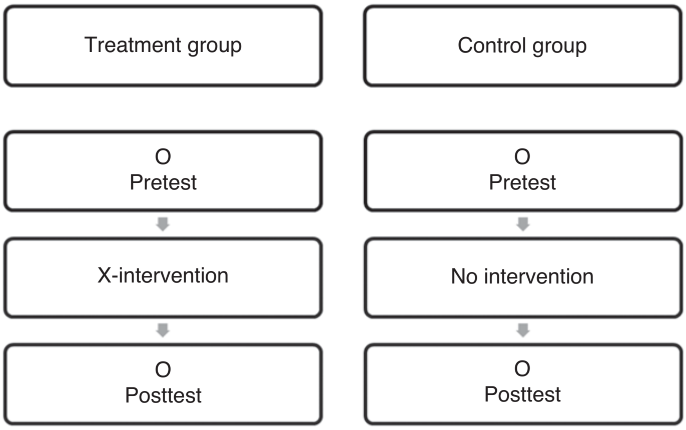


Figure 15.1 *A NECGPP design.*

However, even in NECGPP designs, such intact groups are not always available. Consider, for example, the study by Lee and Lee (2020). Participants were recruited using convenience sampling and were assigned to the experiential/treatment group based on their willingness to engage in the group program; the remaining participants were assigned to the control group. These groups, therefore, were not existing intact groups. Likewise, Noh and Kim (2019) recruited their sample using a convenience method and allocated their groups by time periods. Recruiting the experiential group was conducted over a six-week time period in two separate years, as they chose to use a non-synchronized design to prevent diffusion of treatment. It is very important, therefore, that in such samples, the threats to internal validity are accounted for at the methodology stage of the research.

The series of steps in a NECGPP design are:

(1) Prior to carrying out the research, match characteristics in treatment and control groups.
(2) Carry out a pretest with participants in both the treatment and control groups.
(3) Try to ensure that both the treatment and the control groups experience the same conditions, excluding the intervention.
(4) Carry out a posttest with participants in both treatment and control groups.
(5) Assess changes in the dependent variable between and within groups using specific statistical tests (see data analysis section).

Strengths and Weaknesses of a NECGPP Design

Non-equivalent designs are frequently used in social and behavioral research (Cook & Campbell, 1979; Heppner et al., 1992; Huck & Cormier, 1996; Huck et al., 1974). The NECGGP design allows for a comparison between groups before and after an intervention, and within groups (Heppner, 1999; Heppner et al., 1992). In addition, the simplicity of a NECGPP design ensures that researchers can replicate such methodological approaches, and many argue for the use of such designs (Asher, 1983) especially in educational settings (Denny et al., 2017). Although all researchers endeavor to ensure that measures undertaken in a study are rigorous, carrying out research in social and behavioral science settings may mean that there will be aspects of a study that cannot be controlled.

External and Internal Validity Factors in a NECGPP Design

A NECGPP design can account for some threats to internal validity, such as the uncontrolled threats of sample selection bias (external validity factors), history, maturation, testing effects, complex human variables, Pygmalion effect, and compensation rivalry (Braaten, 1989; Cook & Campbell, 1979; Campbell & Stanley, 1966; Hains & Szyjakowski, 1990; Kush & Cochran, 1993).

Sample Selection

In educational research in a class or intact group, it may not always be possible to carry out random selection of participants. Therefore, the non-random assignment of participants to control and treatment groups can yield groups with different characteristics. Validity in the absence of random selection is identified as a critical element in generalization of findings (Serlin & Lapsley, 1985). Non-random assignment in research tends to show greater bias in results and can severely limit the conclusions that are drawn by the researcher or the generalization of the results to the whole population (Rosenthal & Rosnow, 2008). If convenience or available sampling is used, it cannot be considered representative of the population of interest. However, Heppner et al. (1992, p. 274) refer to the *good enough principle* by which non-random samples can have sufficient characteristics, such that generalization to certain populations is reasonable. In some research studies, generalization – an external validity factor – may not be one of the primary goals of the research.

However, the inclusion of the participant control group adds to the validity of a NECGPP design and could allow findings to be generalized to other settings having population and ecological validity (testing environment) factors that are similar (external validity factors) and that utilize a NECGPP design (Campbell & Stanley, 1966). For example, statistical tests (techniques) can be used for creating reliable comparison group(s) in a NECGPP design. These try to reduce the risk in selection bias and include regression discontinuity design, which has been resurrected in recent years, and a more contemporary approach that is often used is propensity score matching (see Thistlethwaite & Campbell, 1960; White & Sabarwal, 2014).

History

History as an internal threat to validity refers to an external event or exposure that was inadvertently experienced by the control group and that could have an effect on the findings. For example, the control group being exposed to the new teaching and learning approach that was used in the intervention group.

Maturation

During the period of a research study, internal threats occur due to real changes in the environment of participants (e.g., biological and psychological changes; Saks & Allsop, 2007). Using the same example, the treatment group is introduced to a new teaching and learning intervention where the researcher is interested in ascertaining if the learning intervention improved students' end of year exam results. Therefore, one would expect that as both control and treatment participants progress through a three- or four-year degree program, they become better educated – this could have an impact on their learning and development and the outcomes of the research. However, because of the inclusion of a control group in a NECGPP, it would be expected that both groups would have had exposure to similar experiences over the duration of the study.

Testing Effects

An example of a testing effect is familiarity with a questionnaire/instrument, when used pretest and posttest in a NECGPP design, which may enable participants to perform better on the second or subsequent measurements, merely because of their familiarity with the questionnaire (Rosenthal & Rosnow, 2008).

Statistical Regression

Statistical regression to the mean is the tendency for those scoring extremely high or low on a selection measure to score less extreme during subsequent testing (Rosenthal & Rosnow, 2008). For example, in a NECGPP study, you can administer a particular measurement tool at several time points: pretest, midway, and posttest. However, if the time interval phases in a NECGPP are lengthy, statistical regression may not pose a threat to testing on a specific measure. It is noteworthy that statistical regression to the mean results from a selection of subjects based on extreme scores/characteristics, so in a NECGPP study, if groups are matched, regression to the mean for both groups should be about the same.

Complex Human Variables

The influence of several complex human variables requires consideration in a NECGPP. These are not unique to NECGPP designs but are noteworthy, particularly in research where the researcher may be known to the groups. The potential positive effect of the researcher on the experimental group is known as the *halo effect* (Thorndike, 1920). Participants knowing that they are under observation and therefore achieving higher scores in an assessment, for example summative or formative scores in exams, is known as the *Hawthorne effect* (Parsons, 1974). See Chapter 11 in this volume for a more extensive discussion of experimenter effects.

Pygmalion Effect

It is possible that an increase in treatment effect in a research study could be due to the high expectations on the part of the researcher, which can influence the behaviors of participants. This is known as the *Pygmalion effect* (Rosenthal and Jacobson, 1968).

Compensatory Rivalry

Compensatory rivalry or the *John Henry effect* (Barrett & White, 1991) can occur in a control group (being more motivated than participants in the treatment group). In addition, the unintentional exposure of participants in the control group to the treatment condition (e.g., the teaching intervention) may affect the control group results.

External Factors

Despite the possible limitations outlined above, a NECGPP design is frequently viewed as a good option for non-experimental research. However, at the outset of a NECGPP study, many characteristics can be controlled by using selection criteria. The use of NECGPP designs encourages a flexible approach to both the design of the research and to the interpretation of the findings (Robson, 2002). All research designs suffer from threats to validity, and many rival hypotheses exist regarding the findings, even in randomized control designs, when one compares pretest to posttest results (Campbell & Stanley, 1966; Heppner, et al., 1992; Loftin & Madison, 1991; Rosenthal & Rosnow, 2008). Much of the variance in the dependent variable is due to individual differences among participants; the researcher endeavors to reduce the error found in the dependent variable to create a more powerful statistical test (Heppner et al., 1992; Huck & Cormier, 1996; Rosenthal & Rosnow, 2008). This allows the researcher to perform various analyses (e.g., analysis of covariance [ANCOVA]) that may be helpful in making valid inferences (Campbell & Stanley, 1966).

Sampling in a NECGPP Design

Sampling involves the process of selecting representative units of a population for inclusion in a study and ensures the external validity of a study (Heppner et al., 1992; Hoinville & Jowell, 1978). A NECGPP design is commonly employed when random assignment is not possible in practice (Cook & Campbell, 1979). Probability sampling is used when the purpose of the evaluation is to generalize from a sample to the entire population while non-probability sampling is used when it is not possible to use a random probability sample. The researcher can use an available or convenience sample (see other non-probability methods; Saks & Allsop, 2007). Goodwin (1995, p. 109) suggests that non-probability sampling will only result in the findings being extended beyond the research sample "if the relationship studied is a powerful one" and, furthermore, that "it will occur for most subjects within a population regardless of how they were chosen."

Consequently, in deciding on the population in a NECGPP design, the researcher identifies the population descriptors that form the basis for the eligibility criteria (Bell, 1993). The researcher can use a non-random method of sampling with a NECGPP design, described as convenience or available sampling. This sampling frame allows the researcher to select participants that are most readily available and suited to the research question(s) (Denny et al., 2017; Patton, 1990). Non-random sampling is advantageous as it allows the researcher to get samples that otherwise would be unavailable and is particularly suited to a NECGPP design (Heppner et al., 1992). As discussed, the good enough principle – which stipulates that non-random samples can have sufficient characteristics – suggests that generalization to a certain population is reasonable (Heppner et al., 1992). Ideally, the control group is chosen to be as similar as possible to the intervention group (e.g., by matching). For example, in a study by Chiva-Bartoll et al. (2020), existing groups were used; two different years of a sport science and primary school bachelor's degree were recruited to a control and experimental/treatment group. The researchers analyzed the effects of a service-learning program on the subjective happiness, prosocial behavior, and perceptions of professional learning.

Power and Effect Size in a NECGPP Design

In terms of ascertaining sample size estimation for a NECGPP design, the seminal work of Jacob Cohen (Cohen, 1969), in his book entitled *Statistical Power Analysis for the Behavioural Sciences* (revised in 1988; Cohen, 1988), is commonly used. The power of a statistical test is the probability that the test will find a statistically significant effect in a sample size *n*, at a pre-specified level of alpha (α), given that an effect of a particular size exists in the population (Rosenthal & Rosnow, 2008). In a statistical test, alpha is the probability of a Type I error, the probability of rejecting the null hypothesis (H_0, the null hypothesis, is that there is no effect or no relationship between variables) when the null hypothesis is true.

According to Cohen (1962), power is a monotonic function of sample size, and judgments relating to sample size should adhere to conventional standards that will facilitate the performance of power analyses for the most common statistical tests. He stated that (Cohen 1962, p. 153) "[s]ince power is a direct monotonic function of sample size, it is recommended that investigators use larger sample sizes than they customarily do. It is further recommended that research plans be routinely subjected to power analysis, using as conventions the criteria of population effect size . . ." Power analyses are considered increasingly important in social and behavioral sciences and are used to determine the appropriate number of participants to use in a study (Miles, 2003). Power depends not only on sample size, but also on effect size and the chosen significance level of the test (Sokal & Rohlf, 1981). Power is a continuum that varies non-linearly and gradually with sample size (Kazdin &

Bass, 1989). Cohen (1962. p. 147) offers the following conceptual description of the major elements of power:

> The decision- or significance-criterion, α: this is the expression of the researcher's policy with regard to risking the mistaken rejection of a H_0 in the form of a long-term error rate for rejecting when H_0 is true. By stressing that α is a policy, the implication is intended that it pre-exists the gathering of the data and should not be confused with p, the tail area of the statistic derived from the results of the research. The researcher is endeavoring to have enough statistical power so that the null hypothesis can be rejected when some given alternative hypothesis is true.

The researcher is endeavoring to have enough statistical power in a NECGPP design so that the null hypothesis can be rejected when some given hypothesis is true (Cohen, 1973). Cohen suggests that any value can be taken for α, but that the general rule regarding Type I error is to set it as $\alpha = 0.05$. Going back to the earlier example of a teaching strategy intervention using a NECGPP design, $\alpha = 0.05$ is chosen a priori since there is no justification for using a more or less stringent α. The power also measures the chance of detecting an effect size of a known magnitude using the specified experimental design and varies according to the magnitude of the effect specified (Cohen, 1988; Lipsey and Wilson, 1993; see also Denny et al., 2017). An effect size can be measured in two ways: as the standardized difference between two means (most often expressed as *d*) or as the correlation between the independent variable categorization and the individual scores on the dependent variable (i.e., correlation; Rosnow & Rosenthal, 1996). Cohen (1962, p. 146) poses the question: "How large an effect (a difference, correlation coefficient, etc.) in the population do I expect actually exists, or want to be able to detect?" Cohen (1988) refers to three levels of effect size; small = 0.2; medium = 0.5; and large = 0.8 (see Sage Case Study 2 in Denny et al., 2017).

In setting the level of risk, one should have enough power to make $\alpha = \beta$ (Stevens, 2002). β is the probability of a Type II error, the probability of not rejecting H_0 when H_0 is false, that is, β is 1 − power. Employing the traditional $\alpha = 5\%$, and setting $\beta = \alpha$ (Stevens, 2002), would mean a power of 95%. However, getting 95% power in a NECGPP design is not always possible because of the nature of the research setting. If no a priori specific type of research has been carried out in the setting, there are no previous guidelines as to what constitutes an appropriate sample size to use. In this case, a common convention is to try to get at least enough data to have 80% power (Cohen, 1988). Although this is somewhat arbitrary, a power of 80% has become the conventional standard (Heppner et al., 1992). Heppner et al. (1992, p. 278) have theorized that, because statistical significance can be found for *trivial effects*, it is advisable to report the effect size and power in addition to the significance level α, especially in a NECGPP design where the sample size is small.

The magnitude of the effect size between the independent and dependent variable in a NECGPP design is very important. Kirk (2005) discusses the concept of effect size research and suggests that it falls into three categories:

- measures of effect size (standardized mean differences)
- measures of strength of association
- other measures.

Kirk (2005, p. 83) suggests that these measures are used for three purposes: *integrating the results of empirical research studies in meta-analyses, supplementing* information supplied by the null hypothesis significance tests, and determining whether *research is practically significant* (i.e., the usefulness of the results). In terms of representing effectiveness of an intervention, and depending on the research question, researchers may cite gain scores (difference from pretest to posttest scores; a positive score indicates a gain whereas a negative score indicates a decline) and effect sizes (Zimmerman & Williams, 1982). Gain scores were widely used in the intervention field in both comparative and control designs ; however, their use has presented a problem in terms of the interpretation, as other rival explanations for any observed difference could exist (e.g., regression toward the mean; Heppner et al., 1992; Rosenthal, 1984; Kim & Steiner, 2019).

In summary, when determining power in a NECGPP design, the neglect of alpha, sample size, or effect size can have major implications for interpreting research (Kazdin & Bass, 1989). When the researcher is endeavoring to maintain a balance between the risk of a Type I error and the demands of the hypothesis test, an α level of 5% is considered an appropriate value (Gravetter & Wallnau, 2000) and statistical power set at 80%. This is based on the context of the research (e.g., educational setting) using intact groups or an available sampling frame (Denny et al., 2017).

Analysis of Data in a NECGPP Design

The nature of a NECGPP design requires that an appropriate method of data analysis be employed. The main methods of data analysis include the following statistical methods: paired *t*-test; independent two-samples *t*-test; Pearson's correlations, and ANCOVA. The independent two-sample *t*-test is used to compare means from two independent groups of individuals, whereas the paired *t*-test is used to compare means of two sets of observations from the same individuals or from matched pairs of individuals (Brace et al., 2003). The assumptions of the pooled (equal variances) independent two-sample *t*-test are (Bonate, 2000):

- independence – is the data between and within groups independent?
- continuous – is the dependent variable on a continuous scale?
- normality – is the distribution of scores for the dependent variable in the population normal for each level of the independent variable?
- homogeneity of variances – is the variability of the dependent variable in the population similar for each level of the independent variable?

To ascertain if significant differences exist between treatment and control groups in a NECGPP design at pretest, independent two-sample *t*-tests (two-tailed) can be conducted (see Denny et al., 2017). The normality assumption can be verified by means of normal probability plots and more formal tests – Kolmogorov–Smirnov and Shapiro–Wilk. The homogeneity of variances assumption can be examined using box plots and formally using Levene’s test for equality of variances (Pallant, 2006). These assumption checks can verify that normality and homogeneity of variances were met for the dependent variable.

However, a two-sample *t*-test is reliable only if these assumptions are met (Cleveland, 1993). In cases when the homogeneity of variances is not a valid assumption, the Welch–Satterthwaite *t*-test may be applied, as it only assumes normality. The Welch–Satterthwaite *t*-test is an alternative to the pooled independent two-sample *t*-test and is used when the assumption that the two populations have equal variances seems unreasonable. It provides a *t*-statistic that asymptotically (i.e., as the sample sizes become large) approaches a *t*-distribution, allowing an approximate *t*-test to be calculated when the population variances are not equal (Pallant, 2006).

Additionally, researchers using a NECGPP design may be interested in baseline differences between groups, such as differences in age, and this can be investigated using the independent two-sample *t*-test. However, if age, for example, is not normally distributed within each group, a nonparametric test, the Mann–Whitney U-test (two-tailed) can be performed to ascertain if significant differences exist between treatment and control groups at pretest on age.

In the NECGPP design, paired *t*-tests can be conducted to investigate if a significant increase or decrease has occurred in the intervention group between pretest and posttest of full intervention (e.g., testing the effectiveness of a teaching intervention on summative or formative academic results). However, it is accepted that there are many rival hypotheses for any potential change between pre- and posttesting. Performing a paired *t*-test is acceptable, but there are problems interpreting or drawing conclusions.

One-tailed tests can be used if the researcher is interested in whether the obtained value of the statistic falls within one tail of the sampling distribution for that statistic. In contrast, two-tailed tests can be used when the researcher is using a hypothesis that predicts a relationship, but not whether scores increased or decreased.

Taking the same example, separate paired *t*-tests (two-tailed) can also be conducted to investigate if significant changes occurred in either treatment or control groups between pretest and posttest. The assumptions underlying a paired *t*-test are (Bonate, 2000):

- independence – are the data within groups independent?
- continuous - is the dependent variable on a continuous scale?
- normality – is the distribution of differences in scores for the dependent variable in the population normal?

The assumption of normality is assessed using a normal probability plot and more formal tests – Kolmogorov–Smirnov and Shapiro–Wilk. If the differences are not normally distributed, a non-parametric test, such as the Wilcoxon signed-rank test, can be applied.

Another statistical technique used in NECGPP is ANCOVA. ANCOVA involves adjusting the observed dependent variable for the effects of a covariate. It can be viewed as a combination of analysis of variance and regression. ANCOVA was developed by Fisher (1971) to reduce error variance in randomized experiments. It increases the statistical power of hypotheses tests and the precision in estimating effects. Fisher (1971, p. 281) stated that "it combines the advantages and reconciles the requirements of the two very widely applicable procedures known as regression and analysis of variance."

The more frequently used analysis of variance (ANOVA) allows the comparison of several groups, while regression analysis provides a model that relates the dependent variable to the covariate(s). ANOVA involves determining whether the difference between two or more means is statistically significant, while ANCOVA builds one more level of complexity (Kerlinger, 1986).

ANCOVA is a method that compares different groups adjusting for the effect of concomitant or nuisance variables (e.g., age or prescores) using the exemplar already outlined, from summative or formative examination results. Concomitant variables are factors not of direct interest in a study that have an influential effect on the variability of the outcome. With ANCOVA, the differences between the means are examined while also *controlling* for the effects that another variable or variables may have on the dependent variable (Hopkins, 2016). These other variables are typically called covariates. In other words, ANCOVA attempts to remove the effect of the covariate(s) by using a regression equation to measure its influence (Fisher, 1971). For that reason, ANCOVA allows for the removal, from the dependent variable, of any irrelevant or error variance that cannot be predicted from the independent variable (Bonate, 2000). Consequently, by accounting for covariate(s), a more accurate and reliable proportion of variance is obtained – that is, the statistical power is increased (Tabachnick & Fidell, 2001).

The assumptions underlying ANCOVA include the usual ANOVA assumptions:

- The samples are independent (i.e., the data for one group does not depend on the data for the other group, and the data within each group are also independent).
- The data is normally distributed within each group.
- The variability within each group is the same. This permits the computation of one common or pooled estimate of standard variation for all groups. This assumption is often referred to as the homogeneity of variances assumption.

ANCOVA rests on additional assumptions:

- Within each group, the dependent variable has a linear relationship with the covariates (Tabachnick & Fidell, 2001).
- The slope of the regression line for each covariate is the same in each group (parallel-line assumption) and not zero (Tabachnick and Fidell, 2001).

These assumptions can be checked by applying residual diagnostics. This involves determining the predicted and residual values (i.e., the difference between an observed value for the dependent variable under consideration and its predicted value by the estimated ANCOVA model). The normality assumption can be assessed by a normal probability plot of the residuals and more formally performing Kolmogorov–Smirnov and Shapiro–Wilk tests on the residuals. The homogeneity of variances assumption can be verified using a plot of the residuals against the predicted values. The specific ANCOVA assumption of linearity between the dependent variable and covariate(s) within each group can be checked by means of scatter plots. The parallel-line assumption of ANCOVA can be assessed by examining if the slope of the regression line for each covariate is the same in the control and treatment group and not zero (Tabachnick & Fidell, 2001).

If, for example, it was important to establish if age should be considered, using the example cited earlier, one could use age as a covariate in the ANCOVA. To justify its exclusion as a covariate in analyses of scores, separate correlational analyses (i.e., Pearson's correlation) can be conducted for each of the treatment and control groups. If age is non-normally distributed within each group, Spearman's rank order correlations (using two-tailed tests) can be computed and tested for significance. If age shows no significant relationship with posttest of intervention scores, within each group, it is not considered as a covariate. Instead of comparing gain scores between groups, it is recommended that ANCOVA is employed to compare differences on posttest scores using pretest scores as a covariate (i.e., adjust for pre-existing differences between treatment and control groups) in a NECGPP design, if it is assumed that the two research groups (treatment and control) are not equivalent (since the participants had not been randomly assigned to groups).

Effect sizes are commonly used in a NECGPP design. According to Pastor and Kaliski (2007), the most uncomplicated approach is to report the pretest and posttest average scores, with the difference between the averages representing the typical change in raw scores over time. However, these authors note that a disadvantage of this approach is its dependency on the score scale being employed. They cite, as an exemplar, that a typical gain of 5 points appears large on a 20-point scale but negligible on a 100-point scale. For that reason, it is desirable to report standardized measures of change. Currently, standardized measures of an effect are often conveyed using effect sizes and are normally used to capture practical significance using Cohen's *d* or eta-squared (Pastor & Kaliski, 2007).

Significance tests do not simply test the presence or absence of an effect; they are conditional on the effect size – "the degree of departure of the effect from the H_0" (Houle et al., 2005, p. 415). A common effect size is eta-squared (η^2), which indicates the relative magnitude of the differences between means. That is, it describes the "total variance in the dependent variable that is predictable from knowledge of the levels of the independent variable" (Tabachnick & Fidell, 2001, p. 52). A partial eta-squared ($\eta_p{}^2$) statistic can be used in ANCOVA, as it takes the variance attributable to the effect of interest plus the error variance into account (Pallant, 2006; Tabachnick & Fidell, 2001). Guidelines for interpreting eta-squared values suggest that 0.01 indicates a small effect; 0.06, a moderate effect; and 0.14, a large effect (Pallant, 2006).

The final set of analyses that can be performed in a NECGPP design are correlational analyses to examine the associations between changes in dependent variables for the treatment and control groups, using Pearson correlations (two-tailed tests) or Spearman's correlations. For all statistical tests, the significance can be determined using $p < 0.05$, if this was the margin of error accepted.

Additional Comments on Data Analyses

Screening can be performed for outlying data in a NECGPP design, using exploratory data analyses techniques that can look for univariate and multivariate outliers (Tabachnick & Fidell, 2001; Barnett & Lewis, 1994). If outlier scores are identified, which lie outside of the pattern of data, one can observe the distribution of numerical values using histograms and box plots (using the median, and lower and upper quartiles). Then, data collection and transcription can be checked for mistakes. Data analyses in the NECGPP design can be conducted with and without these outliers, and assumption checks can be carried out. If the presence of outliers does not affect the results, the findings from the complete data can be presented (Denny et al., 2017). Multivariate statistical techniques, such as MANCOVA, can also be considered for data analysis in NECGPP designs. MANOVA is a statistical option to test the significance of group differences between groups. However, MANOVA relies on the use of many dependent variables in a NECGPP design.

Hershberger (2005, p. 867) suggests that:

> As ANOVA can be extended to the analysis of covariance (ANCOVA), MANOVA can be extended to testing the equality of group means after their dependence on other variables has been removed by regression. In the multivariate analysis of covariance (MANCOVA), we eliminate the effects of one or more confounding variables (covariates) by regressing the set of dependent variables on them; group differences are then evaluated on the set of residualized means.

Although an alpha level of 5% is increasingly considered the maximum acceptable rate for Type I error, Bordens and Abbott (2007) have suggested that applied research may be evaluated more effectively at a less conservative alpha level. For the ANCOVA parallel-line assumption, to ensure that when a non-parallel line is indicated by means of a significant interaction it is genuine, a stricter level of significance of 1% should be employed. In addition, when determining if age should be used as a covariate, a 1% level of significance should be used.

Exemplar of a NECGPP Design

This section provides two examples of research studies that used NECGPP design.

Example Using Existing Intact Groups

Authors: Chiva-Bartoll et al. (2020).

Study aim: The aim of the study was to analyze the effects of a service-learning program on the subjective happiness, prosocial behavior, and professional learning perceptions of physical education teacher education students as well as to examine the correlations among these variables.

Sampling method: Two existing groups were used from the sport science and primary school bachelor's degree, third- and fifth-year students.

Sample size calculation: Not provided by authors. They used intact groups with 55 in the control group and 49 in the intervention group.

Intervention: The control group were taught using traditional methodologies, while the experimental group were taught using a service-learning model of pedagogy.

Data collection timepoints: Not detailed by authors.

Data analysis: The Kolmogorov–Smirnov test was used to determine the normality of the data. After extracting the mean and standard deviation as descriptive statistics, the *p*-value was calculated using the Wilcoxon test for related samples to identify significant differences. Regarding the pretest and posttest differences between the control and the intervention groups, the Mann–Whitney U-test was used for two independent samples. The effect size was calculated using Cohen's *d* value. The Spearman correlation coefficient was used to determine the relationships between the variables.

Example Where Groups Were Created from a Convenience Sample

Authors: Lee and Lee (2020).

Study aim: The aim of the study was to conduct a group cognitive behavioral program focusing on cognitive processes and behavioral changes to improve the mental health of undergraduate students, in order to identify how the factors of depression, self-esteem, and interpersonal relationships are changed through a NECGPP design.

Sampling method: Participants were recruited through a recruitment advertisement. Because of cultural concerns about the stigma of mental illness and the need to be available for the group program, participants who easily agreed to engage in the group program were first assigned to the experimental group in view of the participation time and grade, and the rest were assigned to the control group. All participants had to meet pre-set inclusion and exclusion criteria.

Sample size calculation: The sample size was calculated (based on a previous study) and it was identified that the minimum sample size required for a *t*-test with $\alpha = 0.05$, power $\beta = 80\%$, and effect size 0.40 was 36 subjects in both groups.

Intervention: The experimental group engaged in a cognitive behavioral group program twice a week for one month.

Data collection timepoints: The pre-survey was measured one week before the program in both the experimental and control groups. The post-survey was conducted immediately after the eighth session.

Data analysis: Descriptive statistics were used to analyze the general characteristics and variables. The chi-square test, Fisher's exact test, and the *t*-test were used to examine the homogeneity in the response variables between the experimental and control groups. To verify the effect of the intervention by time between the experimental group and control group, a repeated measures ANOVA was performed. Two-tailed tests and a 5% significance level were used in all analyses.

Conclusion

The rationale for using a NECGPP design in social and behavioral research has been presented, and we acknowledge that this type of design has limitations. However, its utility and applicability stems from its efficacy in research, where experimental designs are not always suitable and where random selection is neither feasible nor practical. Consideration has been afforded to the important areas in the research process, namely, design, sample selection, power, effect size, and specific methods of data analyses approaches. Finally, we note that having sufficient statistical power (affected by the alpha level, sample size, the use of one-tailed versus two-tailed tests and effect size) is a necessary and important consideration if a researcher is to use a NECGPP design.

Further Reading

Barber, T. X. (1973). Pitfalls in research: Nine investigator and experimenter effects. In R. Travers (ed.). *Second Handbook of Research on Teaching*. Rand McNally.

Cook, D. L. (1967). *The Impact of the Hawthorne Effect in Experimental Design in Educational Research*, Cooperative Research Project, 1967, No. 1757. US Office of Education.

Gephart, W. J. & Antonoplos, D. P. (1969). The effects of expectancy and other research-biasing factors. *The Phi Delta Kappan*, *50*(10) 579–583. https://www.jstor.org/stable/20372478.

Lee, J. (2021). Situation, background, assessment, and recommendation stepwise education program: A quasi-experimental study. *Nurse Education Today*, *100*, 104847. https://doi.org/10.1016/j.nedt.2021.104847

Noh, G. O. & Kim, M. (2021). Effectiveness of assertiveness training, SBAR, and combined SBAR and assertiveness training for nursing students undergoing clinical training: A quasi-experimental study. *Nurse Education Today*, *103*, 104958. https://doi.org/10.1016/j.nedt.2021.104958

Osman, K. & Lee, T. (2014). Impact of interactive multimedia module with pedagogical agents on students' understanding and motivation in the Learning of electrochemistry. *International Journal of Science & Mathematics Education*, *12*(2), 395–421. https://doi.org/10.1007/s10763-013-9407

Rosenthal, R. (1963). On the social psychology of the psychological experiment: The experimenter's hypothesis as unintended determinant of experimental results. *American Science*, *51*, 268–283.

Whitley, E. & Ball, J. (2002). Statistics review 4: Sample size calculations. *Critical Care*, *6*(4), 335–341. https://doi.org/10.1186/cc1521

Yu, F.-Y. & Chen, C.-Y. (2021). Student- versus teacher-generated explanations for answers to online multiple-choice questions: What are the differences? *Computers & Education*, *173*, 104273. https://doi.org/10.1016/j.compedu.2021.104273

References

Asher, H. B. (1983). *Casual Modelling*. SAGE Publications.

Barnett, V. & Lewis, T. (1994). *Outliers in Statistical Data*, 3rd ed. John Wiley & Sons.

Barrett, A. C. & White, D. A. (1991). How John Henry effects confound the measurement of self-esteem in primary prevention programs for drug abuse in middle schools. *Journal of Alcohol and Drug Education*, 36(3), 87–102.

Bell, J. (1993). *Doing Your Own Research Project*. Open University Press.

Benjamin, L. (1988). *A History of Psychology*. McGraw-Hill.

Bonate, P. (2000). *Analysis of Pretest–Posttest Designs*. Chapman & Hall.

Bordens, K. & Abbott, B. (2007). *Research Design and Methods: A Process Approach*. McGrath Hill.

Braaten, L. J. (1989). The effects of person-centred group therapy. *Person Centred Review*, *4*(2), 18.

Brace, N., Kemp, R., & Snelgar, R. (2003). *SPSS for Psychologists. A Guide to Data Analysis using SPSS for Windows*, 2nd ed. Palgrave.

Campbell, D. T. & Stanley, J. C. (1966). *Experimental and Quasi-Experimental Designs for Research*. Rand McNally.

Chiva-Bartoll, O., Montero, P. J. R, Capella-Peris, C., & Salvador-García, C. (2020). Effects of service learning on physical education teacher education students' subjective happiness, prosocial behavior, and professional learning. *Frontiers in Psychology. 11*, 331.

Cleveland, W. S. (1993). *Visualising Data*. Hobart Press.

Cohen, J. (1962). The statistical power of abnormal social psychological research. *Journal of Abnormal and Social Psychology*, *65*(3), 145–153.

Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*. Academic Press.

Cohen, J. (1973). Eta-squared and partial eta-squared statistics in fixed factor ANOVA designs. *Educational and Psychological Measurement*, *33*, 107–112.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Lawrence Erlbaum Associates.

Cook, T. D. & Campbell, D. T. (1979). *Quasi-Experimentation: Design and Analysis for Field Settings*. Rand McNally.

Crotty, M. (2006). *The Foundations of Social Research: Meaning and Perspectives in the Research Process*, 2nd ed. SAGE Publications.

Dawes, M., Davies, P., Gray, A., et al. (2005). *Evidence Based Practice: A Primer for Health Care Professionals*, 2nd ed. Elsevier Churchill Livingstone.

Denny, M., Denieffe, S. & Pajnkihar, M. (2017). *Using a Non-equivalent Control Group Design in Educational Research*. *Research Methods Cases Part 2*. SAGE Publications.

Dickinson, K. P., Johnson, T. R., &. West, R. W. (1987). An analysis of the sensitivity of quasi experimental net impact estimates of CETA programmes. *Evaluation Review*, *11*, 452–472.

Fisher, R. A. (1971). *The Design of Experiments*, 8th ed. Oxford University Press.

Frank, M. G. & Gilovich, T. (1988). The dark side of self- and social perception: Black uniforms and aggression in professional sports. *Journal of Personality and Social Psychology*, *54*(1), 74–85. https://doi.org/10.1037/0022-3514.54.1.74

Goodwin, J. C. (1995). *Research in Psychology: Methods and Design*. John Wiley & Sons.

Gravetter, F. J. & Wallnau, L. B. (2000). *Statistics for the Behavioural Sciences*. Wadsworth/Thomson Learning.

Hains, A. A. & Szyjakowski, M. (1990). A cognitive stress-reduction intervention program for adolescents. *Journal of Counseling Psychology*, *37*(1), 80.

Heppner, P. P. (1999). Extending the tradition of the counseling psychologist by building on strengths. *The Counseling Psychologist*, *27*(1), 59–72. https://doi.org/10.1177/0011000099271005

Heppner, P. P., Kivlighan, D. M., & Wampold, B. E. (1992). *Research Design in Counseling*. Brooks/Cole Publishing Company.

Heppner, P. P., Kivlighan, D. M., & Wampold, B. E (2004). *Research Design in Counseling*, 2nd ed. Brooks/Cole Publishing Company.

Hershberger, S. L. (2005). History of multivariate analysis of variance. In B. Everitt & D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science* (vol. 2, pp. 864–869). John Wiley & Sons.

Hopkins, W. G. (2016). A new view of statistics. Available at: www.sportsci.org/resource/stats/.

Hoinville, J. & Jowell, R. (1978). *Survey Research Practice*. Heinemann.

Houle, T. T., Penzien, D. B., & Houle, C. K. (2005). Statistical power and sample size estimation for headache research: An overview and power calculation tools. *Headache: The Journal of Head and Face Pain*, 45(5), 414–418.

Huck, S. W. & Cormier, W. H. (1996). Principles of research design. In C. Jennison (ed.), *Reading Statistics and Research* (pp. 578–622). 2nd ed. Harper Collins.

Huck, S. W., Cormier, W. H., & Bounds, W. F. (1974). *Reading Statistics and Research*. Harper Collins.

Hunsley, J. & Lee, C.M. (2006). *Introduction to Clinical Psychology*. John Wiley & Sons.

Kazdin, E. & Bass, D. (1989). Power to detect differences between alternative treatments in comparative psychotherapy outcomes research. *Journal of Consulting and Clinical Psychology*, *57*(1), 138–147.

Kerlinger, F. N. (1986). *Foundations of Behavioural Research*. Holt, Reinhart & Winston.

Kim, Y. & Steiner, P. M. (2019). Gain scores revisited: A graphical models perspective. *Sociological Methods & Research*, *50*(3). https://doi.org/10.1177/0049124119826155

Kirk, R. E. (2005). *Handbook of Research in Experimental Psychology*. Blackwell Publishing.

Kush, K. & Cochran, L. (1993). Enhancing a sense of agency through career planning. *Journal of Counseling Psychology*, *40*(4), 434–439.

Lee, S. & Lee, E. (2020). Effects of cognitive behavioral group program for mental health promotion of university students. *International Journal of Environmental Research and Public Health*, *17*(10), 3500.

Lipsey, M. W. & Wilson, D.B. (1993). The efficacy of psychological, educational and behavioural treatment: Conformation from meta-analysis. *American Psychologist*, *48*, 1181–1209.

Loftin, L. & Madison, S. (1991). The extreme dangers of covariance corrections. In B. Thompson (ed.), *Advances in Educational Research: Substantive Findings, Methodological Developments*. JAI Press.

Miles, J. (2003). A framework for power analysis using a structural equation modelling procedure. *BMC Medical Research Methodology*, *3*, 27.

McMillan, J. H. (2000). *Educational Research: Fundamentals for the Consumer*. Addison Wesley Longman.

Murray, T. R. (2003). *Blending Qualitative and Quantitative Methods in Theses and Dissertations*. Corwin Press Inc.

Noh, G. O. & Kim, D. H. (2019). Effectiveness of a self-directed learning program using blended coaching among nursing students in clinical practice: A quasi-experimental research design. *BMC Med Education*, *19*(1), 225.

Parsons, H. M. (1974). What happened at Hawthorn? *Science*, *183*, 93.

Patton, P. Q. (1990). *Qualitative Evaluation and Research Methods*, 2nd ed. SAGE Publications.

Pallant, J. (2006). *SPSS Survival Manual*, 2nd ed. McGrath Hill.

Pastor, D. A. & Kaliski, P. K. (2007). Examining college students' gains in general education. *Research and Practice in Assessment*, *1*(2), 1–20.

Polit, D. F. (2005). *Essentials of Nursing Research: Methods, Appraisal and Utilization*, 6th ed. Lippincott Williams & Wilkins.

Polit, D. F. & Beck, C. T. (2004). *Nursing Research: Principles and Methods*, 7th ed. Lippincott Williams & Wilkins.

Polit, D. F., Beck, C. T., & Hungler, B. P. (2006). *Nursing Research: Methods, Appraisals, and Utilization*, 6th ed. Lippincott Williams & Wilkins.

Robson, C. (2002). *Real World Research: A Resource for Social Scientists and Practitioner-Researchers*, 2nd ed. Blackwell.

Rosenthal, R. (1984). *Meta-analytic Procedures for Social Research*. SAGE Publications.

Rosenthal, R. & Jacobson, L. (1968). *Pygmalion in the Classroom: Teacher Expectation and Pupils' Intellectual Development*. Rinehart and Winston.

Rosenthal, R. & Rosnow, R. L. (2008). *Essentials of Behavioural Research: Method and Data Analysis*, 3rd ed. McGrath Hill.

Rosnow, R. L. & Rosenthal, R. (1996). Computing contrasts, effect sizes and counter nulls on other people's published data: General procedures for research consumers. *Psychological Methods*, *1*, 331–340.

Rubin, D. B. (1979) Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, *74*, 318–328.

Riley, M. W. (1967). *Sociological Research; A Case Approach*. Harcourt Brace and Jovanovich

Saks, M. & Allsop, J. (2007). *Health Research Sampling Methods*. SAGE Publications.

Serlin, R. C. & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist*, *40*, 73–83.

Sokal, R. R. & Rohif, F. J. (1981). *Biometry: The principles and practices of Statistics in Biological Research*. W. H. Freeman and Company.

Stevens, J. (2002). *Applied Multivariate Statistics for the Social Sciences*, 4th ed. Erlbaum.

Tabachnick, B. G. and Fidell, L. S. (2001). *Using Multivariate Statistics*, 4th ed. Harper Collins.

Thistlethwaite, D. & Campbell, D. (1960). Regression–discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, *51* 309–317.

Thorndike, E. L. (1920). A constant error on psychological rating. *Journal of Applied Psychology*, *4*, 25–29.

White, H. & Sabarwal S. (2014). *Quasi-experimental Design and Methods, Methodological Briefs: Impact Evaluation 8*. UNICEF.

Zimmerman, D. W. & Williams, R. H. (1982). Gain scores in research can be highly reliable. *Journal of Educational Measurement*, 19(2), 149–154.

16 Experimental Methods

Thomas F. Denson and Craig A. Anderson

Abstract

This chapter provides an accessible introduction to experimental methods for social and behavioral scientists. We cover the process of experimentation from generating hypotheses through to statistical analyses. The chapter discusses classical issues (e.g., experimental design, selecting appropriate samples) but also more recent developments that have attracted the attention of experimental researchers. These issues include replication, preregistration, online samples, and power analyses. We also discuss the strengths and weaknesses of experimental methods. We conclude by noting that, for many research questions, experimental methods provide the strongest test of hypothesized causal relationships. Furthermore, well-designed experiments can elicit the same mental processes as in the real world; this typically makes them generalizable to new people and real-life situations.

Keywords: Experiments; Experimental Methods; Experimental Design; Generalizability; Replication; Sample Characteristics; Statistical Power

Introduction

Social and behavioral scientists are tasked with uncovering truths about a vast range of phenomena related to human behavior. To do so, scientists test hypotheses derived from theories (see Chapter 1 in this volume). The use of experimental methods provides a strong means of hypothesis testing. The aim of the experiment is deceptively simple: to quantitatively determine the causal effect of the independent variable(s) on the dependent variable(s). The criteria for conducting a good experiment are strict and usually require some degree of material and personnel resources and willing volunteers to participate. When experiments are properly conducted, no other scientific tool has the ability to confer such a high level of confidence in causal relationships between variables. Perhaps not surprisingly, results from experiments form a large part of the knowledge base in the social and behavioral sciences as well as every other scientific discipline.

What Do We Mean by an Experiment?

Our primary focus here is what is known as the true experiment. The true experiment is characterized by one or more carefully crafted experimental manipulations, which are known as the independent variable(s). Part of the manipulation process is to include a control condition that is identical to the experimental condition(s) in all aspects except the manipulated independent variables of interest. The outcome of interest is known as the dependent variable. For example, dependent variables include an observation such as helping behavior, self-reported attitudes toward an ethnic outgroup, or physiological measures (e.g., heart rate).

In a true experiment (hereafter to be called "experiment"), the study's unit of analysis often is each person who participates in the study ("participants"). In other studies, the unit of analysis might be various groups of people, such as existing decision-making committees or temporarily created discussion groups. The statistical analyses required for groups or dyads is somewhat more complicated than the analyses for individuals. For simplicity, we focus on individual participants as the units of analysis, but the concepts in this chapter apply to other units of analysis. In experimental studies, each participant is randomly assigned to different levels of the independent variable.

The simplest experiment consists of two levels of the independent variable (e.g., playing a violent video game or a non-violent video game for 20 minutes). The dependent variable could be anything of theoretical and/or practical interest (e.g., aggressive thinking or amount of salivary cortisol; for both dependent measures, see Gentile et al., 2017). In this particular study, children played a violent game or a non-violent game (i.e., the independent variable). Aggressive thinking was the dependent measure. Participants filled in missing letters for numerous word fragments, some of which can be completed to form an aggression-related or aggression-unrelated word. For example, "ki_ _" can become either "kiss" or "kill". The proportion of all completed words that use the aggressive option is a measure of the extent to which that person was recently thinking aggressive thoughts.

We can think of each child's aggressive thinking score as being made up of several components. First, any measurement (e.g., meters, grams, aggressive thinking) has two main components, a true component (by definition unknown) and an error component. The measured score for each person (e.g., aggressive-thinking score) is usually designated as Y_i, the true score as T_i, and the error component as E_i. The error component itself can be further broken down into two subcomponents: random and systematic. Random error (or noise) is inevitable, but too much can make detecting a true effect more difficult. The other, called systematic error, is even more serious because it can lead to false conclusions about the hypothesized effect of the independent variable on the dependent variable. For example, the measured score of each individual may be systematically related to characteristics of the person (e.g., how often their parents used physical punishment during childhood).

Let's consider what this means for the Gentile et al. (2017) study. If the researchers had allowed each child to choose which game they would play for 20 minutes, then it is quite likely that their choices would have been affected by some unmeasured (and uncontrolled) third variable (e.g., parental punishment practices, gaming history, personality traits). In other words, the researchers should not have much confidence that the result (i.e., higher aggressive thinking by those who played the violent game) was caused by what type of game they played. Maybe the kids who chose the violent game varied systematically in some important way from those who chose the non-violent game.

By randomly assigning participants to play either one game or the other, such as by flipping a coin, the likelihood of the two groups differing in some important way, such as aggression-related personality traits, decreases dramatically. Thus, these researchers converted systematic error into random error, thereby increasing the probability that any difference in aggressive thinking (or cortisol) between the two game groups was caused by the 20 minutes of violent vs. non-violent game play. In short, random assignment to different experimental conditions allows researchers to draw strong causal inferences about the effect of the independent variable on the dependent variable. Because of this feature, the experiment distinguishes itself from other research designs. Of course, experiments have weaknesses, and other research designs (e.g., correlational, longitudinal, naturalist/quasi-experiments) also contribute to the ultimate scientific conclusions concerning what variables cause what effects and under what conditions. The following sections outline the various stages involved in designing an experiment and the decisions that need to be made along the way.

Generating Hypotheses

The first step is to create a testable hypothesis about how two or more variables are related (see Chapter 3 in this volume). Sometimes this begins by observing or reading about some interesting event or attending a research conference. At other times, the initial idea comes from carefully working out the implications of some existing theory. In either case, the next step should involve reading the relevant scientific literature to find out what already is known, what is unknown, and what theories seem most relevant (see Chapter 4 in this volume). Once the research team has thoroughly examined the literature, they can start to think about stating one or more hypotheses that they would like to test.

The hypothesis states the expected relationship between an independent variable and the dependent variable. For instance, hundreds of experiments have tested the hypothesis that brief exposure to violent entertainment media (e.g., violent television shows, films, video games) compared to non-violent media elicits aggressive behavior in players (for video games, see Anderson et al., 2010; Greitemeyer & Mügge, 2014). In this type of experiment, half of participants play a violent game (experimental condition) and half play a non-violent game (control condition). Afterward, participants are given an opportunity to behave aggressively. Aggression is measured on a quantitative scale (e.g., number of electric shocks delivered to another participant) that can be compared between conditions using a statistical test.

In order to test our primary hypothesis (also called *alternative hypothesis*; H_1), we must also define the *null hypothesis* (H_0; see Chapter 22 in this volume). This null hypothesis is what one should expect if the independent variable has no effect on the dependent variable. In this case, the null hypothesis is that the mean number of electric shocks delivered by the violent video game group will be equal to the mean number of shocks delivered by the non-violent video game group. The alternative hypothesis would state that the mean number of shocks would be greater in the violent video game condition than the non-violent video game condition. One key aspect of a true experiment is that the alternative hypothesis must be falsifiable, meaning that it can be proven wrong. Specifying a falsifiable hypothesis is a prerequisite for scientific inquiry and guides the decisions made in developing the experimental conditions, dependent measures, sample population, and statistical analyses.

Experimental and Control Conditions

A hypothesis specifies *what* idea will be tested, whereas deciding on the parameters of the experimental and control groups determines *how*. One of the most difficult tasks of conducting experiments is determining how to select a good control group. Ideally, the experimental and control groups are identical except for the experimental manipulation. If the control group is too similar to the experimental group, the effect of the independent variable cannot be differentiated from the control condition. If the control group is so dissimilar that it contains none of the features of the experimental condition, results will likely support the experimenter's hypothesis; however, the researchers will be unable to determine which of the different features produced the effect.

For example, a researcher might wish to test the extent to which drinking alcohol will make people aggressive toward sexual minorities. The alternative hypothesis is that alcohol intoxication will elicit more aggression than sobriety. With this hypothesis in mind, the researcher must determine what participants will drink in the experimental and control conditions. In one such experiment, Parrott and Lisco (2015) asked 320 heterosexual men, some of whom reported being prejudiced toward sexual minorities, to consume an alcoholic drink in the experimental condition and a non-alcoholic drink in the control condition. A common control condition in such studies is to have participants consume a placebo drink. The advantage of the placebo is that participants in both groups will think they have consumed alcohol. This is important because some people believe alcohol will make them aggressive, and therefore having a control group that knows that they didn't consume alcohol results in a potential confound (see Chapter 13 in this volume). Thus, the similarity between the two levels of the independent variable is greater in a placebo-controlled experiment than doing the same experiment but with a no-drink control condition. However, placebos are rarely, if ever, consumed in the real world; therefore, a no-drink condition more closely resembles real life than a placebo. The authors found that alcohol intoxication (relative to the non-alcoholic drink condition) increased aggression toward a homosexual man among the prejudiced heterosexual participants but not among those low in prejudice.

Choosing a control group is often challenging because the independent variables themselves cannot always be directly observed. Therefore, one needs to first define the independent and dependent variables at the conceptual level. This process is typically accomplished by extensively consulting the research literature and defining the independent and dependent variables of interest in accordance with existing theoretical perspectives. Next, the researcher must operationalize the variables of interest (i.e., determine how each will be measured and/or manipulated). The complexity of going from theory to operationalization (also called "empirical realization") is illustrated in Figure 16.1, which shows the various interpretation steps often needed to get from basic theory to selecting how to create two levels of the independent variable. This same kind to translation is needed for the dependent variable as well.

This process of defining and operationalizing variables can be tricky. Take, for example, terror management theory (Burke et al., 2010). This theory posits that fear of death motivates a substantial amount of human behavior, including the need to seek solace in one's cultural worldview and behave as an upstanding cultural citizen. Because one's culture will exist beyond their death, being a part of something that exists longer than oneself alleviates some of the death anxiety. But how does one operationalize the fear of death and turn it into a valid manipulation with an appropriate control group? One solution is to temporarily increase mortality salience by having participants write about what will physically happen to them when they

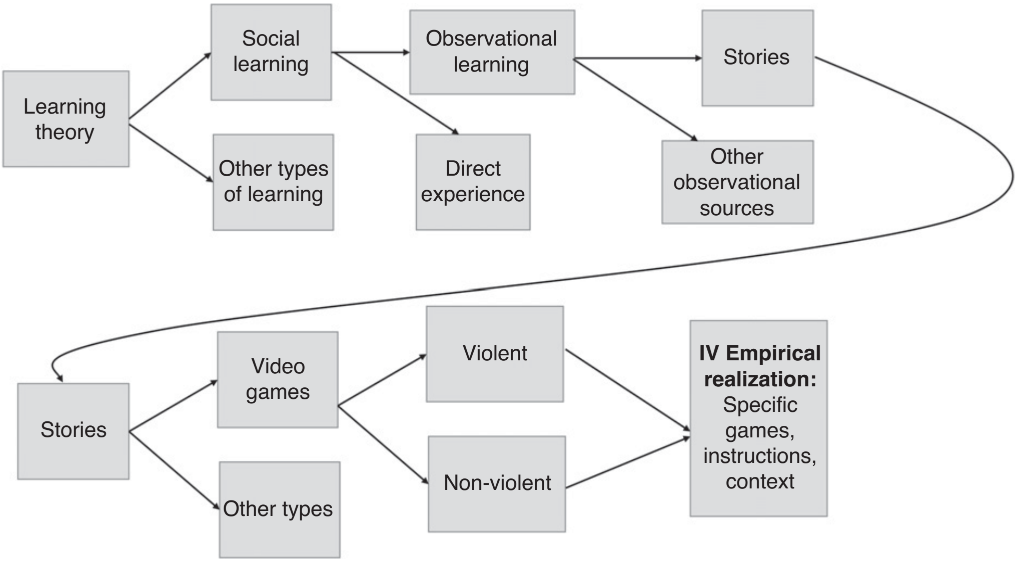


Figure 16.1 *Illustration of multiple translation levels from learning theory to empirical realization of the independent variable: experimental manipulation of video game violence (Prot & Anderson, 2013).*

die. Creating a control condition that matches the mortality salience manipulation on all aspects except the fear of death is tricky. Most experiments settled on writing about dental pain for the control condition (Burke et al., 2010). This example shows the difficulty of choosing a well-matched control condition.

Choosing the Sample

After creating the experiment and obtaining ethical approval (see Chapter 2 in this volume), researchers recruit participants. Several decisions are made when selecting an appropriate sample, and these can influence the research quality and the extent to which the results generalize to other people. The first decision is to choose a sample that is appropriate for the research question. For example, if one wishes to compare the effect of two types of peer interaction (e.g., mobile phone games versus book club) on life satisfaction during retirement, a sample of undergraduates will not suffice. The best approach to sample selection might be to randomly sample from the population of interest (e.g., all retired people in Australia) and then randomly assign participants to conditions (e.g., phone games or book club). Because this sample would be representative of the population of interest, such an approach would allow generalization to the population (e.g., retired Australians). However, this approach is unwieldly and highly unlikely to occur in practice.

Recruitment

Finding the appropriate samples can often be difficult (see Chapter 6 in this volume). In the social and behavioral sciences, the most common sampling technique is a convenience sample because they are easily recruited by university-based researchers (i.e., comprised of participants who are ready, willing, and available to participate in research). When convenience samples are not very representative of the general population to which the scholar wishes to generalize, the researcher cannot be confident that the same results would occur in other populations. In recent decades, social and behavioral scientists increasingly try to access more appropriate samples, including convenience samples online. Two hugely popular sources of online participants are Mechanical Turk (www.mturk.com) and Prolific (app.prolific.co). These fee-based services offer rapid access to large samples. One study sought to determine if differences existed between three convenience samples: standard undergraduates, undergraduates who use Mechanical Turk, and Mechanical Turk users who were not undergraduates (Weigold & Weigold, 2021). They found differences between the three groups on all the variables they investigated, including demographics, time to study completion, attention to the study, personality, social desirability, need for cognition, values, and attitudes. Thus, researchers should be aware of these differences and consider how they might affect their experiment's outcomes.

Culture

The social and behavioral sciences have been criticized for their reliance on college undergraduate convenience samples. In a highly influential paper, Henrich et al. (2010) took this criticism one step further and noted that most psychological research had been conducted in Western, educated, industrialized, rich, and democratic (so-called "WEIRD") nations. Scientists often seek to discover universals of human behavior that apply to all people. Thus, researchers sometimes assume that findings from a WEIRD convenience sample of college undergraduates will generalize to other populations – populations that differ by country, language, age, race/ethnicity, poverty/wealth, personality and traits. Henrich et al. (2010, p. 29) noted ". . . that research articles routinely assume that their results are broadly representative, rarely adding even a cautionary footnote on how far their findings can be generalized." This overgeneralization problem has become increasingly obvious to many scholars, leading to greater use of diverse samples and to greater caution about overgeneralizing results (see Pettigrew, 2021).

Random Assignment Revisited

There are other common difficulties in conducting experiments with people. As noted earlier, there are two types of error in all measurements – random error and systematic error. Random error is by definition "random," (i.e., not correlated with the independent variables). Too much random error makes it difficult to detect small true effects, just as conducting a hearing test in a loud natural environment (e.g., at a baseball game) would result in erroneous results in prescribing hearing aids. This cost to science and society is that real true effects are missed because of the random error. Systematic error is even worse because it leads to incorrect conclusions about what causes what, thereby harming development of good theory and application. As noted at the outset, random assignment to different experimental conditions is key to removing many sources of systematic error. We now highlight a few additional sources of systematic error in human experimental research.

Experimenter and Participant Bias

In many experiments, the person who administers the experimental manipulation (commonly called the experimenter) is very much aware of the hypothesis and of the experimental condition in which any given participant is placed. This knowledge can lead the experimenter to treat the participants in systematically biased ways (see Chapter 11 in this volume). For example, one might *unintentionally* treat participants who are in the stress condition in a more brusque manner than those assigned to a no-stress condition. If the brusqueness of the experimenter has an impact on the outcome variable (e.g., production of stress hormones), the results of the experiment are contaminated by the experimenter's unintentional but systematically different treatment of participants in different experimental conditions. Such experimenter biases can work either in favor of the main hypothesis, or against it. In both cases, the results are invalid.

Several common solutions reduce this problem. One is to standardize the interactions between experimenters and participants. This standardization can be done by creating a very specific script for experimenters to follow. Still, people often give unintentional non-verbal cues that can systematically vary by condition. Such cues might be avoided by using video or audio/video recordings for all interactions. Another effective technique is to design the experiment in a way that keeps the experimenter from knowing which experimental condition participants are in. In other words, the experimenter is “blind” to the condition participants have been assigned.

A related problem concerns the knowledge of the research participant. People bring beliefs, expectations, and emotions with them wherever they go. Knowledge of the purpose of the study and/or of the specific condition that they will be in (e.g., alcohol vs. non-alcoholic drink) can lead people to behave in artificial ways, sometimes intentionally so. One solution is to keep participants ignorant of these details (to the extent that it is ethically feasible to do so). Thus, a “double-blind” procedure is one in which both the participant and experimenter are unaware of which condition the participant is in.

In addition to using “blinded” procedures, a good research team can reduce participant bias by creating a really good cover story that disguises the true purpose of the study. Another important technique is to make the experimental situation very involving and impactful. This can reduce both the time and the effort that the participant has to generate hypotheses about how they are “supposed” to behave and can decrease artificial responding.

Types of Experimental Designs

Between-Subjects Design

Experimental design specifies how many groups and time points will be included in an experiment. In the simplest possible experiment, participants are randomly assigned to either one experimental condition or one control condition, and one dependent variable is measured once. This two-group design can be expanded out to a potentially vast number of other experimental conditions to compare with the control condition. However, in practice, more than eight groups are rarely seen in behavioral research. This type of design – one or more experimental groups and a control group – is known as a *one-way design*. There is one factor with several “levels” that differ from each other qualitatively (as with different induced emotions) and/or quantitatively (as with different doses of alcohol).

The use of a one-way design is illustrated in an experiment on emotion regulation by Kalokerinos et al. (2015). Participants recruited from Mechanical Turk were

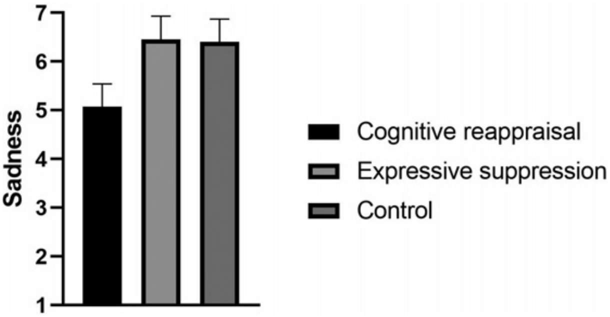


Figure 16.2 *Data from a one-way between subjects experiment on emotion regulation (Kalokerinos et al., 2015). The independent variable was emotion regulation strategy with three levels. The dependent measure was self-reported sadness after watching a sad film clip.*

randomly assigned to view a sad film clip from the *Lion King* in one of three ways: while viewing the film from a detached, unemotional perspective (cognitive reappraisal), while suppressing all overt emotional responses (expressive suppression), or as they normally would (control). Thus, the independent variable constituted two forms of emotion regulation and a control group, thereby creating a one-way between-subjects design with three levels. Participants in the cognitive reappraisal condition reported feeling less sad than participants in the expressive suppression and control groups (Figure 16.2).

The one-way design can be expanded to more complex models that include two or more independent variables that are sometimes called factors. A design with two or more factors is known as a *factorial experimental design*. This type of design is typically used to see if the effects of one independent variable on the dependent variable differ as a function of the second independent variable. For example, a researcher might test the hypothesis that a pain manipulation would increase aggression against a competing partner, but only when the partner insulted them first. This is a 2 (pain induction: yes vs. no) × 2 (verbal insult: yes vs. no) between-subjects factorial design. There would be four conditions: pain and verbal insult, pain and no insult, no pain and verbal insult, no pain and no verbal insult.

Using a 2 × 2 factorial design in a different context, Riva et al. (2015) tested the possibility that mild electrical stimulation to a part of the brain that is implicated in emotion regulation would help people who were socially excluded behave less aggressively. They manipulated brain stimulation (active versus placebo) and whether people were excluded or included by other participants in a ball-tossing game. Thus, the design was a 2 (brain stimulation: active, placebo) × 2 (social exclusion: included, excluded) between-subjects factorial design. After these experimental manipulations were done, participants were given the opportunity to make one of the other players consume as much hot sauce as the participant selected for them. Participants had been informed that the other players disliked spicy foods, and they would have to consume the entire amount of hot sauce that the participant allocated. This dependent variable is commonly used to measure aggression in laboratory studies (Lieberman et al., 1999). In reality, the other "players" were computer agents programmed by the researchers. Results showed that, for participants who were included in the ball-toss game, the brain stimulation factor had no effect on hot sauce allocation, presumably because no emotion regulation was required in this benign situation. However, among the excluded participants, those who were given the active brain stimulation were less aggressive than participants who were given the placebo stimulation. Thus, the effect of brain stimulation on aggression depended upon, or was *moderated by*, the factor of social exclusion. Brain stimulation only "worked" for participants who were socially excluded.

The above examples illustrate what are called between-subjects designs. In such designs, each participant only experiences one of the experimental conditions. In the 2 (brain stimulation: active, placebo) × 2 (social exclusion: included, excluded) between-subjects factorial design just described (Riva et al., 2015), participants were randomly assigned to the active stimulation *or* the sham stimulation *and* the social inclusion condition *or* the exclusion condition. Thus, approximately 25% of participants were randomly assigned to one of the four groups, but they only experienced one level of each factor. Because there are four combinations, there will be four groups. Similarly, a 2 × 3 between-subjects design would have six groups and a 2 × 4 would have eight groups.

Within-Subjects Design

A within-subjects design is used in experiments in which the research team wants the participants to experience all of the conditions. In our emotion regulation example, if participants watched the *Lion King* film clip and were instructed to complete all three levels – reappraisal, suppression, and control, this design would be a one-way within-subjects design. In within-subjects designs, the dependent variable (in this case, self-reported sadness) is measured during or after each condition manipulation; hence, these designs are also known as "repeated measures" designs. Within-subjects designs can be one-way or factorial. If we added another factor to the film experiment, such that all participants viewed the sad clip along with a happy clip, we would have a 2 (film type: sad, happy) × 3 (emotion regulation: suppression, reappraisal, control) within-subjects design. In this hypothetical experiment, all participants would view both clips and complete all three emotion regulation conditions and their sadness and happiness would be measured six times.

Within-subjects designs are often a good choice because each participant acts as their own control/comparison condition of sorts. Because the same participant completes all of the conditions, the researcher can usually detect changes on the dependent variable with good sensitivity. Another reason to use within-subjects designs is when participants are difficult to recruit or expensive. For instance, most studies of brain activity with functional magnetic resonance imaging (fMRI) use within-subjects designs, at least partially due to expense. In one

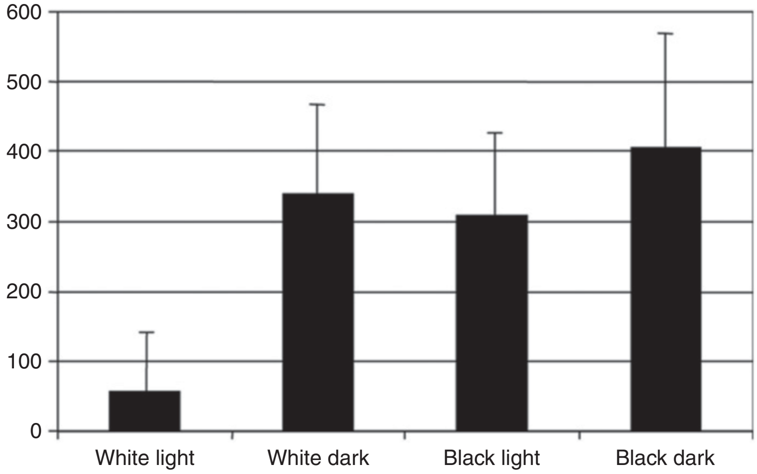


Figure 16.3 *Data from a within-subjects fMRI experiment on amygdala responses to ethnicity (White, Black) and skin tone (light, dark).*

such study, Ronquillo et al. (2007) wanted to test the hypothesis that skin tone would influence the previous finding that Black faces elicit greater responses in the amygdala – a part of the brain that plays a critical role in emotion and threat – than White faces. During scanning, participants were exposed to a 2 (White, Black) × 2 (light-skinned, dark-skinned) within-subjects design in which participants viewed all four combinations of faces. Both types of Black faces and dark-skinned White faces elicited greater amygdala activation than the light-skinned White faces (Figure 16.3).

Mixed Designs and Counterbalancing

Between-subjects and within-subjects designs can be combined such that one (or more) independent variable is between-subjects and one (or more) independent variable is within-subjects. This type of design is known as a mixed design. In one such example, Denson et al. (2014) sought to determine how using cognitive reappraisal during a stressor would influence cortisol output. For the stressor, all participants gave a five-minute speech to two ostensible experts on communication ability. For the between-subjects manipulation, participants were randomly assigned to the cognitive reappraisal condition (e.g., think of the performance in a detached, impersonal manner) or the control condition. The within-subjects variable was the time of cortisol assessment. Cortisol was assessed three times – at baseline, after the stressor, and after a recovery period. Thus, the design was a 2 (cognitive reappraisal, control) × 3 (time: baseline, post-stressor, recovery) mixed design. Results suggested that the extra effort of using cognitive reappraisal augmented cortisol responses to stress (Figure 16.4).

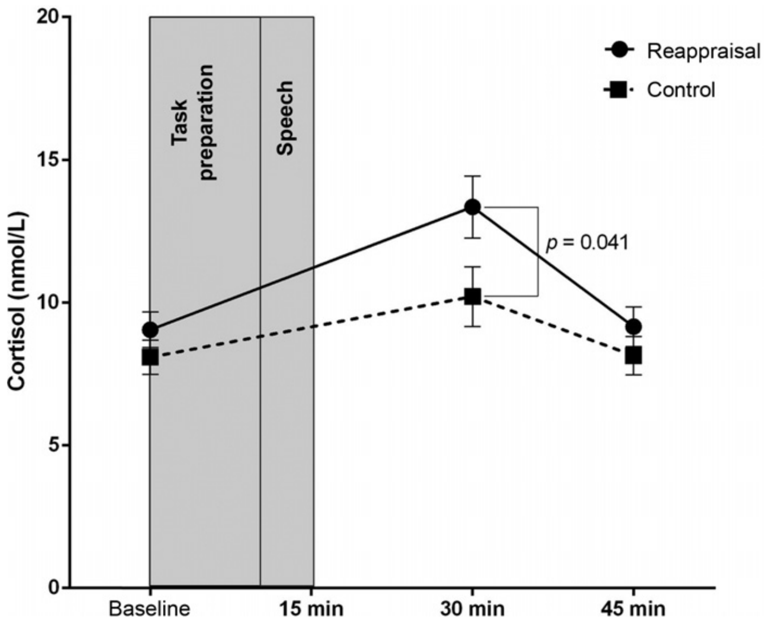


Figure 16.4 *Data from a mixed design experiment on cognitive reappraisal and cortisol responses to a speech stressor (Denson et al., 2014). The between-subjects independent variable was cognitive reappraisal (versus control). The within-subjects dependent measure was salivary cortisol at three time points.*

Mixed and within-subjects designs may elicit concerns about order effects. In our *Lion King* example, suppose the research team was concerned that participants might feel more happiness if they viewed the happy clip first and the sad clip afterward. In other words, the order in which participants are exposed to the levels of the independent variables may affect how participants respond to the dependent variables. This effect is known as an *order effect* and can introduce unwanted variability (systematic error) in the responses on the dependent variable.

The solution to unwanted order effects is *counterbalancing*. In the *Lion King* example, the researcher could randomly assign half of the participants to view the sad clip first and the other half to view the happy clip first. The research team might also be concerned that the order of the emotion regulation instructions could affect the dependent variables. To assess this potential confound, researchers could use all possible orders (e.g., sad clip–reappraisal, sad clip–suppression, sad clip–control, happy clip–control, happy clip–reappraisal, etc.). Because the experiment now contains all possible orders of the independent variable, the research team can be confident that any observed results were not caused by order effects. The systematic error introduced by order effects now becomes random error. Researchers may choose to counterbalance one or more independent variables, depending on the research questions and concerns about order effects. In a *fully counterbalanced design*, all independent variables are counterbalanced; this ensures that all possible orders are part of the experiment.

Quasi-experimental Designs

Many questions that researchers ask cannot be answered with pure experimental methods for practical, ethical, and other reasons. Sometimes, it is impossible or unethical to manipulate a variable. For instance, studies that test the effects of gender on a dependent measure cannot be a true experiment because we cannot randomly assign participants to a gender. Similarly, scientists cannot randomly assign high-school students to be in a "carries weapons" versus "doesn't carry weapons" to school conditions. When a true experiment cannot be used, or is less than ideal for testing hypotheses, researchers may prefer the quasi-experiment (see Chapter 14 in this volume) or other non-experimental methods (see Chapters 13 & 15 in this volume). For instance, Yuan et al. (2020) wished to test the hypothesis that parents of children who were hospitalized during the COVID-19 pandemic (but not for COVID-19 infections) would show more symptoms of anxiety and depression than parents of children who were hospitalized at a time other than during the pandemic. The research team hypothesized that being in a crowded hospital environment during the pandemic versus not during the pandemic would cause parents to become anxious about their children or themselves becoming infected.

The primary limitation of quasi-experimental designs from a true experimentalist's point of view is the lack of random assignment. If this were a true experiment, parents of children in need of hospitalization would have to be randomly assigned to experience a pandemic or not. One can immediately see the unethical and absurd impossibility of conducting this experiment. Because parents cannot be randomly assigned to a pandemic or non-pandemic condition, we cannot be certain that observed differences in anxiety and depression between the two groups was caused by pandemic-inspired concerns. For instance, the heightened general anxiety in the parents of children hospitalized during the pandemic may be due to unemployment or worry about job security or drinking more alcohol, all of which might be correlated with poverty, race, and ethnicity – each of which is correlated with the likelihood of getting infected. In sum, when true experimentation is not possible, quasi-experiments may be informative, providing that the research team and readers are aware of the limitations and discuss them accordingly.

Sample Size and Statistical Power

To determine how many people should participate, researchers should use statistical power analyses to ensure that the sample is large enough to obtain a definitive result. If too few people participate, the researchers will be unlikely to have enough data to sufficiently test their hypothesis. If too many people participate, the researchers will have unnecessarily wasted participants' time and their own time and resources. Experiments that involve any risk for participants also are required to include samples as small as is scientifically reasonable. For these reasons, collecting data from too few or too many participants may be considered unethical.

To determine the required number of participants for an experiment, research teams typically conduct a *power analysis* (see Chapter 6 in this volume). Statistical power is the likelihood of detecting an effect if there truly is one. The desired power is usually set at 0.80 in the social and behavioral sciences (Cohen, 2013). This means that this level of power will provide an 80% chance of detecting a true effect and a 20% chance of failing to detect a true effect. Failure to detect a true effect is known as a Type II error. The chance of making a Type I error (i.e., a false positive) is usually kept lower than making a Type II error. Concluding an effect is true when it is not (Type I) is usually considered a more egregious error than failing to find a true effect (Type II); the latter retains the status quo before the study was conducted.

An a priori *power analysis* is conducted prior to running the experiment and used to estimate the appropriate number of participants. This analysis uses three interdependent parameters: alpha level, the magnitude of the expected effect, and the desired level of statistical power. The alpha level (i.e., the chance of obtaining a statistically significant result when there is in fact no true effect) is usually set at $\alpha = 0.05$. When set to 0.05, the statistical test is considered significant if the *p*-value is less than 0.05. However, the researchers must also accept that there is a 5% chance of concluding that a result is significant when the significant result is due to chance.

Estimating the size of the expected effect is the most difficult of the three parameters. If the experiment is a replication attempt, the expected effect size can be obtained from previous experiments or a meta-analysis in the same domain. However, in many cases, the research team will be testing a novel hypothesis. In this instance, the research team may consult the literature for studies or meta-analyses in a similar area. For instance, Gable et al. (2015) wanted to test the novel hypothesis that anger relative to a neutral emotional state would narrow the scope of cognitive processing. In the absence of an existing effect size, the research team may have consulted the literature for studies that tested the effects of anger on other forms of cognitive processing (e.g., Moons & Mackie, 2007). They could then use these previous effect sizes to intelligently inform their own effect size estimate.

The expected effect size greatly influences the number of participants needed to obtain the desired alpha and power parameters. All things being equal, to detect a significant result, small true effects will require larger samples to detect than large true effects. There are many different, interchangeable effect size metrics to choose from. In experimental settings, the most common are Cohen's *d*, Hedge's *g*, and eta-squared (η^2). Although relatively arbitrary, values of 0.20, 0.50, and 0.80 are considered small, medium, and large effects, respectively, for Cohen's *d* and Hedge's *g*. Small, medium, and large conventions for eta-squared are 0.01, 0.06, and 0.14. These values are considered arbitrary in part because they do not take into account variability in effect sizes across fields of study. A revised set of benchmarks was proposed by Richard et al. (2003). They found that the average effect size in 18 areas of social psychology was about $d = 0.43$, or eta-squared = 0.04. Thus, it can be argued that, for some fields, the original standards for calling an effect small, medium, or large may be too large.

It is important to keep in mind that even very small effect sizes can have disproportionate practical consequences (Anderson et al., 2003; Prentice & Miller, 1992; Rosenthal, 1990). This is especially true when: (a) the effect accumulates over time (e.g., when individuals in the real word are repeatedly exposed to the same risk/protective factor); (b) the risk/protective factor is (or could be) present in a very large population; and (c) when the outcome variable is very important (e.g., life or death). In the medical domain, for example, experimental studies (called "randomized controlled" trails) of the efficacy of new treatments have been stopped early because the treatment proves so effective that further delays in giving the treatment to the placebo condition participants would be unethical. Some such examples had very small effect sizes, according to the benchmarks described earlier. Commonly cited examples include having heart attack survivors take small daily doses of aspirin, using propranolol in heart attack cases, giving organ transplant patients cyclosporine to reduce rejection, and using the drug AZT to treat HIV/AIDS (Rosenthal, 1990). In these cases, the Cohen's *d* effect sizes were about 0.07, 0.08, 0.39, and 0.47, respectively. Note that the largest of these critical breakthroughs in medical science are essentially the same size as the average effect size in social psychology, and the video game violence effect on aggression (Anderson et al., 2010; Richard et al., 2003; Rosenthal, 1990).

Generalizability

A principal aim of the experimentalist is to discover phenomena that extend beyond the laboratory to new people at different times. Behavior in the laboratory is of little interest if that is the only place where it occurs. Extension beyond the experiment is known as generalizability. Two related dichotomies encapsulate how scientists think about generalizability: internal versus external validity and mundane versus experimental realism.

If the independent variable produces a change in the dependent variable, and there is no reason to think that the effect was caused by an uncontrolled variable, the experiment is considered to have high internal validity (McDermott, 2011). Well-designed experiments are considered high in internal validity because procedures, such as random assignment, rule out alternative explanations. Thus, the researcher can be confident that the manipulation caused the observed outcome. External validity refers to the extent to which studies demonstrate replicability across time, situations, people, and operationalizations of the independent and dependent variables. External validity is desirable because it shows that the relationship between the independent and dependent variables stretches outside of the experimental setting.

One common criticism levied at experimental methods is that experiments do not closely resemble real life. The implication is that if the experiment does not closely model real life the results cannot generalize to other people outside those who participated in the experiment. Two forms of realism are applicable to experimental settings. *Mundane* realism refers to the extent to which the context of the experiment resembles the real world. For instance, some alcohol researchers have created “bars” in their laboratories that closely resemble bars in the real world (Bernstein & Wood, 2017). This experimental context is considered high in mundane realism. *Experimental* realism refers to the extent to which psychological processes induced in the experimental context adequately capture the theoretically important underlying processes and variables. That is, participants become so psychologically immersed in the experimental context that they think, feel, and behave as they would in similar real-world contexts. Experiments are often, but not always, low in mundane realism, but well-designed experiments can be very high in experimental realism.

An illustration of mundane and experimental realism is provided by Blake et al. (2020). A long-standing debate in this literature was the extent to which some heterosexual women wear sexualized clothing to feel good about themselves or to please men. In accordance with newer feminism approaches, the researchers tested the hypothesis that beautification would increase assertiveness in undergraduate women. In the first one-way between-subjects experiment, there were two conditions. All participants were asked to bring a change of clothes to the laboratory. In the control condition, participants were asked to change into clothing that they would wear around the house while hanging out with friends. In the beautification condition, participants were asked to change into clothes that they would wear on a “hot date” and had access to their make-up and hair accessories. Participants in the beautification condition scored more assertively on self-report and implicit measures of assertiveness than women in the control group. This experiment had some features that contributed to mundane realism (e.g., using own clothes); however, it is probably very uncommon that someone would get ready for a hot date in the laboratory. In the second experiment, the researchers asked participants to beautify (or not) at home instead of in the laboratory. Thus, Experiment 2 had greater mundane realism than Experiment 1, as participants were at home with their own clothing, beauty aids, etc. Both experiments were reasonably high in experimental realism. Interestingly, the beautification manipulation influenced the assertiveness outcomes regardless of the levels of mundane realism. In sum, as long as experiments are high in experimental realism and internal validity, results can be generalized to new populations even when the experimental setting only superficially resembles the real world. Indeed, internal validity is a prerequisite for external validity and generalization.

Replication

Replication occurs when a scientist conducts an experiment with the intention of copying a previous experiment. There are two types of replication experiments. A *direct replication* occurs when a scientist uses the identical methods from one experiment with a new sample drawn from the same population. A *conceptual replication* is similar, albeit the materials for the independent and/or dependent variables are not identical with the original studies, but instead share a conceptual similarity. Other contextual features, such as the sample population, may differ as well. For example, if the original study tested the effects of violent video games on hostile thinking, a direct replication would use the same video games as the manipulation and the same dependent measure of hostile thinking (e.g., a questionnaire). A conceptual replication might use different violent and non-violent video games, but a conceptually similar dependent measure of hostile thinking and/or different sample population.

The aim of direct replication is to test the robustness of the original effect; the aim of the conceptual replication is to see if the effect can be generalized beyond the specific stimuli and context of the original experiment. Replication is critical to building a knowledge base in every scientific discipline. Discovering how broadly or narrowly an effect applies is critical to theory testing, development, and change. The social and behavioral sciences have a long history of self-examination concerning the validity and reproducibility of major findings. Cohen's (1962) scholarship showing just how poorly powered hypothesis tests are in peer-reviewed psychology journals is one such example; it led (eventually) to stricter standards concerning power, sample size, and effect sizes.

A similar "crisis of confidence" emerged among psychologists in the late 1960s and the 1970s concerning questions about whether laboratory-based studies generalized to the "real" world and whether cultural changes over time makes discovery of "laws of human nature" impossible. This self-examination led to new research testing the generalizability of key laboratory paradigms, including comparisons of memory studies conducted in the field and in the laboratory and similar comparisons in other domains (e.g., leadership, authoritarianism, aggression, depression, and goals). Overall, the finding of numerous such studies is that well-conducted laboratory and field studies usually find the same basic effects (Anderson et al., 1999). It also led to thoughtful explications of when findings should converge, when they should not, and when external validity is irrelevant (e.g., Banaji & Crowder, 1989; Berkowitz & Donnerstein, 1982; Mook, 1983).

The current psychology "replication crisis" was largely triggered by a rather imperfect attempt at testing replicability (Open Science Collaboration, 2015). Almost 300 researchers conducted direct replications of 100 studies (not all were experiments) that had been published in top social and cognitive psychology journals. Though the reported results were eye-opening (47% replication rate), this large-scale project and others sparked a debate about the aims of replication and what constitutes a good replication (e.g., Gilbert et al., 2016). Many scientists were accused of questionable research practices, which threatened reputational damage and elicited indignant retorts. However, recent discussions recognize that one failed replication or even several does not discount whether an effect exists. For example, Edlund et al. (2022) provide a cogent discussion about what information replications can and cannot provide. They also present methods for improving replication quality such as having multiple laboratories work on an effect at the same time and pairing up researchers who are theoretically opposed to conduct a replication (i.e., a so-called "adversarial collaboration").

Despite the overly dire conclusions of the current "replication crisis," it did inspire improvements in scientific practice, just as previous "crises" had done. For example, failed replications can be made publicly available online (e.g., Open Science Framework), journals often require all data and materials to be made available online, new and more uniform statistical methods and reporting guidelines have emerged, and there appears to be greater respect by authors and editors for multi-scholar and multi-country/culture research. So, what makes a good replication?

Conceptual Independent and Dependent Variables and Their Instantiation

In our view, the social and behavioral sciences rarely conduct a true *direct* replication experiment. In its most stringent form, a direct replication requires using the exact same materials and procedures on a *new* sample of participants from the *exact same population* as the original. Even if one did exactly replicate the stimuli and procedures, and sample, from the same population (e.g., Iowa State University undergraduates in the psychology participant pool), recent historical events might have changed the way that population now thinks about the stimuli being presented (e.g., consider the massive effects of the September 11, 2001 terrorist attack in the United States on attitudes toward violence; Carnagey & Anderson, 2007).

One can relax these strict standards a bit, of course, but the researchers must remain cognizant of potential problems. For example, one of us had the privilege of using undergraduate students in his research from several very different US universities; my team discovered that participants at elite US universities have considerably better verbal skills than those at large, not-so-elite universities. This difference required a change in items used to measure self-reported affect. This vocabulary problem might not be a major problem if the focus was on a physiological variable, such as salivary cortisol, rather that verbal reports. Where exactly to draw the line between direct versus conceptual replication is not always clear.

Thus, when planning a study, it might be wise to consider all replications as conceptual. When in this mindset, it becomes easier to think about the planned study as occurring in a particular historical and cultural context. This is necessary to make good decisions about whether or not to use the exact same materials (e.g., US stereotypes about Black people) with a sample of Romanian participants. It might be more appropriate to change the stimuli to reflect local stereotypes about an important minority group, such as the Roma. This contextual thinking (Pettigrew, 2021) is needed to determine the most appropriate independent variable manipulation and dependent variable measure (Figure 16.1). The bottom line is that the researcher needs to determine whether the study is designed to test general theoretical propositions (e.g., provocation effects on anger are larger when the provocateur is a member of a disliked outgroup minority than a liked in-group non-minority) or a much narrower hypothesis.

Preregistration

Preregistration generally is one means of improving research practices and improving replicability. Preregistration involves placing a detailed plan of a proposed experiment on an online repository (e.g., the Open Science Framework, Dryad, Github). Doing so is helpful in at least three ways. First, it makes the research team think more clearly about their own study methods and intended data analyses. Second, it reduces the likelihood that the team will engage in inappropriate flexible research practices. Third, if alternative analyses are deemed necessary, the team will need to acknowledge and justify the changes in a public manner. Sometimes such changes are reasonable, so they should not be prohibited, but they should be made public.

Limitations of Experimental Research

If true experiments are the gold standard method of establishing causality, why bother doing other types of studies, such as cross-sectional or longitudinal correlation studies? There are several answers, most of which boil down to the fact that very often an experimental study cannot be done, usually for either ethical or practical reasons. For example, imagine that you have good theoretical reason to believe that experiencing negative racial stereotyping, prejudice, and discrimination on a daily basis is one *cause* of adult substance abuse. An experiment could be designed in which very young children (at birth or even prenatally) are randomly assigned to spend their first 20 years of life in a highly racist environment, a moderately racist environment, or a non-racist environment. At the end of 20 years, you measure substance abuse. Such a study cannot be executed on both ethical and practical grounds. It is unethical to intentionally expose people to a long-term social environment that is expected to yield significant harm to the participants. Obviously, it also is impossible on practical grounds.

In general, ethical considerations prohibit social/behavioral scientists from subjecting humans to experimental conditions that are expected to produce major long-term harm to those participants who have been randomly assigned to those conditions (see Chapter 2 in this volume). But if the “harm” is short-term only (i.e., the harmful effects are expected to dissipate fairly quickly), can experiments be done? Generally, the current answer across most disciplines is yes, as long as the risks and benefits are carefully and accurately explained to participants or to the person responsible for making such decisions for the potential participant (e.g., in the case of children).

Another way to address the ethical problem of harmful effects is to eliminate the harmful exposure condition from the experiment and instead test the effect of removing the causal risk factor. This is done by randomly assigning some participants to a condition in which the hypothesized risk variable is removed or reduced from the participant’s social environment and assigning the other participants to a no-intervention control condition. For example, we can’t ethically assign some children to grow up in conditions that increase children’s exposure to violent entertainment media (e.g., television, video games), because past research shows that there is long-term harm caused by such exposure. However, we can randomly assign children to an intervention designed to reduce their exposure to violent media or to a non-intervention control condition. This allows a clean test of the causal hypothesis that exposure to violent media increases aggressive behavior (Krahé & Busching, 2015).

How do researchers know which social factors are likely to yield harmful vs. beneficial effects? Mostly, this comes from previous correlational, longitudinal, and other types of studies – studies that are based on well-constructed theories. Interestingly, as the evidential base of a particular theory domain gets stronger, as the need for large-scale experimental studies to confirm/disconfirm specific hypotheses about real-world consequences grows, some of the practical limitations on such studies also become less severe. Once it became clear that levels of lead in children's environments were *associated* with lower IQ and other harmful effects, it came possible to conduct large-scale intervention experiments to confirm (or disconfirm) the *causal* hypothesis.

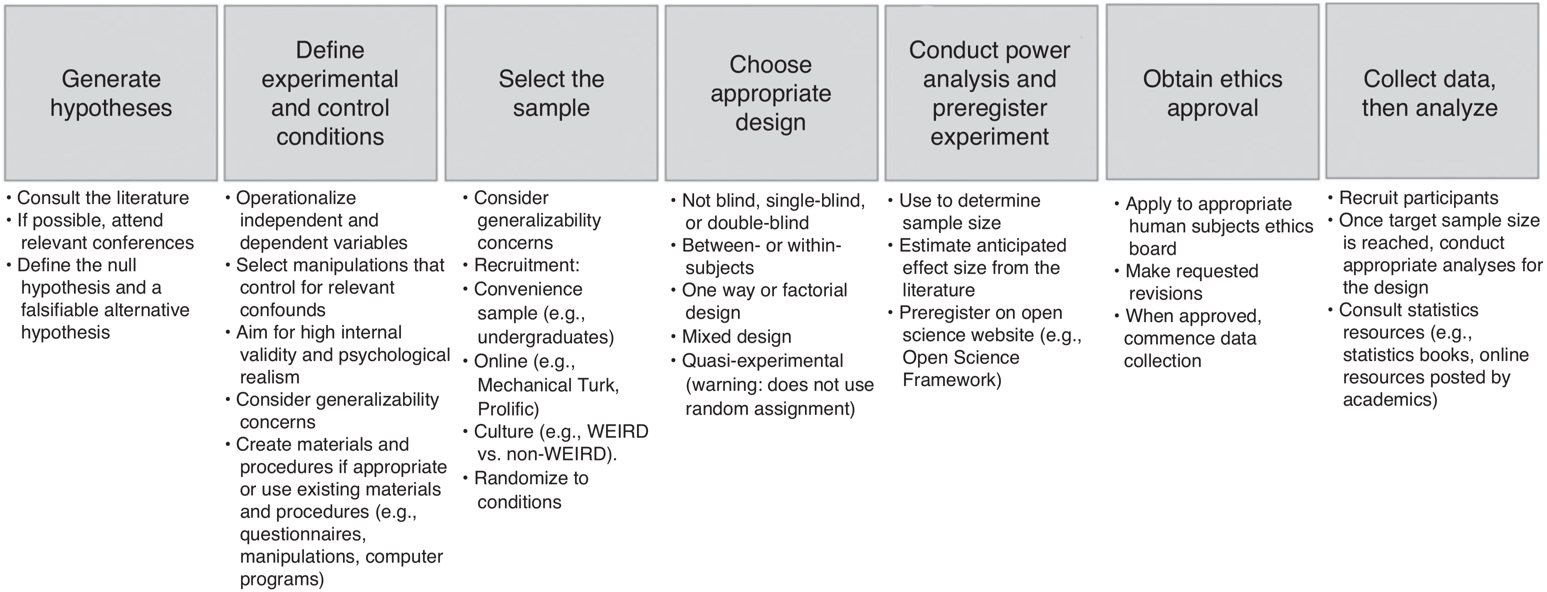
In addition to ethical and practical limitations, there are several social psychological phenomena that may adversely impact the internal validity of laboratory experiments. Research with human participants is inherently a social process; thus, the research team should be aware of these processes and try to mitigate their influence as best they can. Nichols and Edlund (2015) provide a detailed account of several processes. One is participant crosstalk in which former participants discuss the research with future participants. Crosstalk threatens the validity of the findings because some participants are naïve whereas others are not. Lack of naïveté can lead to biased responding. One effective solution to avoid spreading of information about the experiment involves asking participants to sign a statement in which they promise to keep their experience in the experiment confidential.

A second phenomenon is demand characteristics – "the totality of cues which convey an experimental hypothesis to the subject" (Orne, 1962, p. 779). Demand characteristics can guide participants to behave in a manner that is consistent (or sometimes inconsistent) with hypotheses (see Bender et al., 2013, for an example of how gamers may intentionally sabotage results in video game studies). The threat to internal validity is apparent in that participants' behaviors can no longer be ascribed solely to the manipulation. Expectancy effects can also threaten internal validity. These effects occur when the experimenter verbally or non-verbally rewards participants in a manner that is consistent with the experimenter's desired outcomes. When it is feasible to do so, the research should ensure that experimenters are blind to the condition and hypothesis.

Some research paradigms require deceiving participants about the research aims. Participants who are already aware (e.g., through crosstalk) or become aware during the experiment are said to be suspicious. For various reasons, many participants are reluctant to admit suspicion. Nichols and Edlund (2015) provide suggestions to counteract this hesitancy and thereby identify participants who were suspicious, such as ensuring that participants will not be penalized for divulging suspicion, increasing rapport and identification, and highlighting the importance of the research for society. Accounting for all of these confounding variables as best as possible can greatly enhance the internal validity of laboratory experiments.

Conclusion

This chapter provides a brief introduction to the process of conducting experimental research. Figure 16.5 presents a prototypical timeline of the process of experimentation. When conducted properly, the true experiment is the strongest method of inferring causal relations between two variables. Although not all research



Generate hypotheses

- Consult the literature
- If possible, attend relevant conferences
- Define the null hypothesis and a falsifiable alternative hypothesis

Define experimental and control conditions

- Operationalize independent and dependent variables
- Select manipulations that control for relevant confounds
- Aim for high internal validity and psychological realism
- Consider generalizability concerns
- Create materials and procedures if appropriate or use existing materials and procedures (e.g., questionnaires, manipulations, computer programs)

Select the sample

- Consider generalizability concerns
- Recruitment:
- Convenience sample (e.g., undergraduates)
- Online (e.g., Mechanical Turk, Prolific)
- Culture (e.g., WEIRD vs. non-WEIRD).
- Randomize to conditions

Choose appropriate design

- Not blind, single-blind, or double-blind
- Between- or within-subjects
- One way or factorial design
- Mixed design
- Quasi-experimental (warning: does not use random assignment)

Conduct power analysis and preregister experiment

- Use to determine sample size
- Estimate anticipated effect size from the literature
- Preregister on open science website (e.g., Open Science Framework)

Obtain ethics approval

- Apply to appropriate human subjects ethics board
- Make requested revisions
- When approved, commence data collection

Collect data, then analyze

- Recruit participants
- Once target sample size is reached, conduct appropriate analyses for the design
- Consult statistics resources (e.g., statistics books, online resources posted by academics)

Figure 16.5 *An overview of the experimental research process.*

questions lend themselves to experimental designs, when feasible to do so, we suggest implementing experimental methods. We hope that the overview presented in this chapter will prove helpful to those who wish to initiate themselves into the exciting world of experimentation in the social and behavioral sciences.

References

Anderson, C. A., Lindsay, J. J., & Bushman, B. J. (1999). Research in the psychological laboratory: Truth or triviality? *Current Directions in Psychological Science*, *8*, 3–9.

Anderson, C. A., Berkowitz, L., & Donnerstein, E. (2003). The influence of media violence on youth. *Psychological Science in the Public Interest*, *4*, 81–110. https://doi.org/10.1111/j.1529-1006.2003.pspi_1433.x

Anderson, C. A., Shibuya, A., Ihori, N., et al. (2010). Violent video game effects on aggression, empathy, and prosocial behavior in eastern and western countries: A meta-analytic review. *Psychological Bulletin*, *136*(2), 151–173.

Banaji, M. R. & Crowder, R. G. (1989). The bankruptcy of everyday memory. *American Psychologist*, *44*, 1185–1193.

Bender, J., Rothmund, T., & Gollwitzer, M. (2013). Biased estimation of violent video game effects on aggression: Contributing factors and boundary conditions. *Societies*, *3*, 383–398. https://doi.org/10.3390/soc3040383

Berkowitz, L. & Donnerstein, E. (1982). External validity is more than skin deep: Some answers to criticisms of laboratory experiments. *American Psychologist*, *37*(3), 245–257. https://doi.org/10.1037/0003-066X.37.3.245

Bernstein, M. H. & Wood, M. D. (2017). Effect of anticipatory stress on placebo alcohol consumption in a bar laboratory. *The American Journal of Drug and Alcohol Abuse*, *43*(1), 95–102.

Blake, K. R., Brooks, R., Arthur, L. C., & Denson, T. F. (2020). In the context of romantic attraction, beautification can increase assertiveness in women. *PloS One*, *15*(3), e0229162.

Burke, B. L., Martens, A., & Faucher, E. H. (2010). Two decades of terror management theory: A meta-analysis of mortality salience research. *Personality and Social Psychology Review*, *14*(2), 155–195.

Carnagey, N. L. & Anderson, C. A. (2007). Changes in attitudes towards war and violence after September 11, 2001, *Aggressive Behavior*, 33, 118–129.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65(3), 145–153.

Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences*. Academic Press.

Denson, T. F., Creswell, J. D., Terides, M. D., & Blundell, K. (2014). Cognitive reappraisal increases neuroendocrine reactivity to acute social stress and physical pain. *Psychoneuroendocrinology*, *49*, 69–78.

Edlund, J. E., Cuccolo, K., Irgens, M. S., Wagge, J. R., & Zlokovich, M. S. (2022). Saving science through replication studies. *Perspectives on Psychological Science*, *17*(1), 216–225.

Gable, P. A., Poole, B. D., & Harmon-Jones, E. (2015). Anger perceptually and conceptually narrows cognitive scope. *Journal of Personality and Social Psychology*, *109*(1), 163–174.

Gentile, D. A., Bender, P. K., & Anderson, C. A. (2017). Violent video game effects on salivary cortisol, arousal, and aggressive thoughts in children. *Computers in Human Behavior*, *70*, 39–43. http://dx.doi.org/10.1016/j.chb.2016.12.045

Gilbert, D. T., King, G., Pettigrew, S. & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science." *Science*, *351*(6277), 1037. http://dx.doi.org/10.1126/science.aad7243.

Greitemeyer, T. & Mügge, D. O. (2014). Video games do affect social outcomes: A meta-analytic review of the effects of violent and prosocial video game play. *Personality and Social Psychology Bulletin*, *40*(5), 578–589.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466* (7302), 29.

Kalokerinos, E. K., Greenaway, K. H., & Denson, T. F. (2015). Reappraisal but not suppression downregulates the experience of positive and negative emotion. *Emotion*, *15* (3), 271–275.

Krahé, B. & Busching, R. (2015). Breaking the vicious cycle of media violence use and aggression: A test of intervention effects over 30 months. *Psychology of Violence*, *5* (2), 217–226.

Lieberman, J. D., Solomon, S., Greenberg, J., & McGregor, H. A. (1999). A hot new way to measure aggression: Hot sauce allocation. *Aggressive Behavior: Official Journal of the International Society for Research on Aggression*, *25*(5), 331–348.

McDermott, R. (2011). Internal and external validity. In J. N. Druckman, D. P. Greene, J. H. Kuklinski, & A. Lupia (eds.), *Cambridge Handbook of Experimental Political Science* (pp. 27–40). Cambridge University Press.

Moons, W. G. & Mackie, D. M. (2007). Thinking straight while seeing red: The influence of anger on information processing. *Personality and Social Psychology Bulletin*, *33* (5), 706–720.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist*, *38*, 379–387.

Nichols, A. L. & Edlund, J. E. (2015). Practicing what we preach (and sometimes study): Methodological issues in experimental laboratory research. *Review of General Psychology*, *19*(2), 191–202.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, 943. https://doi.org/10.1126/science.aac4716

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, *17*(11), 776–783.

Parrott, D. J. & Lisco, C. G. (2015). Effects of alcohol and sexual prejudice on aggression toward sexual minorities. *Psychology of Violence*, *5*(3), 256–265.

Pettigrew, T. F. (2021). *Contextual Social Psychology: Reanalyzing Prejudice, Voting, and Intergroup Contact*. American Psychological Association.

Prentice, D. A. & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin*, *112*(1), 160–164. https://doi.org/10.1037/0033-2909.112.1.160

Prot, S. & Anderson, C. A. (2013). Research methods, design, and statistics in media psychology. In K. Dill (ed.), *The Oxford Handbook of Media Psychology* (pp. 109–136). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195398809.013.0007

Richard, F. D., Bond, C. F., Jr., & Stokes-Zoota, J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*(4), 331–363. http://dx.doi.org/10.1037/1089-2680.7.4.331

Riva, P., Romero Lauro, L. J., DeWall, C. N., Chester, D. S., & Bushman, B. J. (2015). Reducing aggressive responses to social exclusion using transcranial direct current stimulation. *Social Cognitive and Affective Neuroscience*, *10*(3), 352–356.

Ronquillo, J., Denson, T. F., Lickel, B., et al. (2007). The effects of skin tone on race-related amygdala activity: An fMRI investigation. *Social Cognitive and Affective Neuroscience*, *2*(1), 39–44.

Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist*, *45*(6), 775–777. https://doi.org/10.1037/0003-066X.45.6.775

Weigold, A. & Weigold, I. K. (2021). Traditional and modern convenience samples: An investigation of college student, Mechanical Turk, and Mechanical Turk college student samples. *Social Science Computer Review*. https://doi.org/10.1177/08944393211006847

Yuan, R., Xu, Q. H., Xia, C. C., et al. (2020). Psychological status of parents of hospitalized children during the COVID-19 epidemic in China. *Psychiatry Research*, 288, 112953.

17 Longitudinal Research: A World to Explore

Elisabetta Ruspini

Abstract

This chapter describes some of the issues to be considered when dealing with longitudinal data. Longitudinal data can be defined as data gathered on a set of units over multiple time periods. Longitudinal data can be collected either prospectively or retrospectively, and data can be either qualitative or quantitative. Different ways of deriving repeated observations generate the three main types of longitudinal design: repeated cross-sectional surveys, panel surveys, and retrospective surveys. The world of longitudinal research is thus very heterogeneous. This chapter provides both a summary of advantages and disadvantages of each longitudinal design and some guidelines for authors and researchers.

Keywords: Life Course, Longitudinal Designs, Longitudinal Research, Panel Data, Social Change

Introduction

The aim of this chapter is to provide guidance for researchers interested in approaching the world of longitudinal research. Longitudinal data are, today, a necessity but also a challenge. On the one hand, they are an indispensable tool for the analysis of change at both the micro and macro level. By allowing researchers to follow life courses over time, they can provide important findings about contextual influences on people’s lives and help reconcile theories about social change – developed at the macro-sociological level – with the changing life-course patterns of individuals. On the other hand, they are a challenge because information on the same participants/units is obtained repeatedly over time – longitudinal studies multiply information and can be complex, demanding, and expensive. Moreover, longitudinal research can take many forms. Thus, researchers trying to pursue longitudinal research face a number of challenges. Challenges include how to incorporate temporal issues into theories, how to best design the longitudinal study, how to implement it, and how to analyze and compare different types of longitudinal data (Ployhard & Ward, 2011).

Hence, it is becoming important to both encourage wider use of longitudinal research and to facilitate the exchange of information between those who have already worked and reasoned “longitudinally,” those who would like to do so but are not sure how, and those who are wary of the consequences of approaching and dealing with dynamic data. Within this context, the aim of this chapter is to offer some guidelines and provide ideas to anyone who wishes to carry out longitudinal research. The chapter contains several examples to enable the reader to experience the richness of the world of longitudinal research, as well as its complexity. It is divided into three sections: Traces of History, The Many Faces of Longitudinal Research, and Longitudinal Research: Benefits and Challenges.

Traces of History

Longitudinal research, in the social and behavioral sciences, has a relatively short history (Voelkle & Adolf, 2015). Until very recently, longitudinal data collection and analysis was particularly uncommon in the social sciences, especially in sociology, and mainly only seen in the health sciences. Starting from the eighteenth century, studies designed to gather data about the dynamics of individual phenomena at multiple points in time were mostly found in the fields of medicine, psychology, and anthropometry (Nesselroade & Baltes, 1979; Wall & Williams, 1970). The longitudinal method was used to delineate developmental patterns and etiological relations in the physical growth, personality development, and physiological development of children (Sontag, 1971). However, according to Rajulton (2001), it was not until the 1920s that we find significant longitudinal studies on developmental sequences. One key example is the monumental work undertaken by Lewis M. Terman of Stanford University to study the life histories of gifted children. A thousand gifted children were followed over their life course and gifted adults also were studied backward to the period of childhood, using both prospective and retrospective methods (Terman et al., 1925, 1929, 1930).

Cohort studies – which typically recruit and follow participants who share a common characteristic, experience, or a common event in a selected period (e.g., birth, graduation, or marriage) – were implemented widely between the late 1940s and 1960s principally in the USA and the UK. They were carried out to address pressing public health concerns – the causes of heart disease, the consequences of smoking and risk of lung cancer, and the effects of radiation exposure (Samet & Muñoz, 1998). These studies, by comparing the disease experience of people born at different periods, can be considered some of the most important tools for epidemiological investigation (Doll, 2001; Giroux, 2011). One well-known example is the Framingham Heart Study (FHS), a long-term, ongoing cardiovascular cohort study of residents of the city of Framingham (Massachusetts). It was launched in 1948 with the goal to investigate the epidemiology and risk factors for cardiovascular disease (CVD). The FHS has followed CVD development in three generations of participants. It began with the recruitment of the original cohort: 5,209 people (2,873 women and 2,336 men) between the ages of 30 and 62. In 1971, the study enrolled a second generation – 5,124 of the original participants' adult children and their spouses – and in 2002 a third generation was included in the study, the grandchildren of the original cohort (Caruana et al., 2015; Samet & Muñoz, 1998; Tsao & Vasan, 2015). Another landmark cohort study, the investigation of the atomic bomb survivors in Hiroshima and Nagasaki, addressed the consequences of radiation exposure. It was initiated by the Atomic Bomb Casualty Commission (ABCC) in the 1950s. In 1975, the ABCC was replaced by the Radiation Effects Research Foundation. This study has become one of the principal sources of evidence on the cancer risks of acute radiation exposure (Samet & Munoz, 1998).

Longitudinal research in the forms of panel surveys – and specifically household panel studies that trace individuals and their households over time by gathering information about them at regular intervals – has flourished since the 1970s and 1980s (Menard, 2002; Ruspini, 2002). Longitudinal research on prospective data was initially developed in the USA, where the first household panel in history was launched in 1968 – the Panel Study of Income Dynamics (PSID). In Europe, the German Socio-Economic Panel (SOEP) was set up in 1984 and the British Household Panel Survey (BHPS) began in 1991. These are among the longest running household panel surveys, and were directly inspired by the American PSID. These early, well-known examples have had a significant influence on international research by inspiring and influencing other longitudinal studies, including the Swiss Household Panel (SHP), China Family Panel Studies (CFPS), and the Household, Income and Labour Dynamics (HILDA) survey – a household-based panel that follows the lives of more than 17,000 Australians each year.

The Many Faces of Longitudinal Research

Longitudinal is a rather broad term that implies the notion of repeated measurements (van der Kamp & Bijleveld, 1998). Longitudinal research refers to the collection, analysis, and interpretation of data gathered at multiple points in time, both forward (into the future) and backward (into the past), and both quantitatively and qualitatively. In a longitudinal study, the same set of units (people, families, households, firms, etc.) is followed across two or more periods; the participants in a typical longitudinal study are asked to provide information regarding the issues of interest on a number of separate occasions.

Longitudinal studies vary significantly in terms of their length, duration, sequence of interviews (i.e., number of "waves"), time interval between successive waves, methods of data collection (e.g., pen-and-paper personal interview, computer-assisted personal interview, computer-assisted telephone interview, computer-assisted web interview), size and complexity. Surveys are also increasingly being administered in multiple modes (Longhi & Nandi, 2015; Venkatesh & Vitalari, 1991). For example, the BHPS data collection has been based on pen-and-paper personal interview techniques and face-to-face interviews since its inception in 1991, but wave 9 of the BHPS went into the field using computer-assisted personal interview methodology for the first time in September 1999 (Laurie, 2003). The SHP initially conducted interviews exclusively by telephone but, since 2010, this panel has offered alternative modes to people who were unwilling to respond by telephone (face-to-face and web-based interviews).

The inclusion of time in data design and collection generates different types of studies. Indeed, different longitudinal designs exist (see, for example, Blossfeld & Rohwer, 2013; Dale & Davies, 1994; Laurie, 2013; Rafferty et al., 2015; Ruspini, 2002; Taris, 2000). The most commonly used longitudinal designs are repeated cross-sectional studies, prospective longitudinal studies (cohort and panel surveys), and retrospective longitudinal studies. Each design is briefly described below.

Cross-Sectional Studies

The cross-sectional design is a type of research that studies a cross-section of the population at a specific point in time. The term "cross section" indicates a wide sample of people of different ages, ethnic groups, educational attainments, religious beliefs, and so on. Most often, cross-sectional data are data for micro units – individuals, households, companies, etc. However, cross-sectional data can also be collected on macro units (e.g., municipalities, counties, or even countries). Cross-sectional surveys provide a snapshot of the characteristics of the target population and what is happening at a given time point, offering an instant, but static, "photograph" of the processes being studied. It is, thus, impossible to infer causality. Because of this, cross-sectional surveys are occasionally repeated twice or more. Repeated cross-sectional data are created when a survey is administered at successive time points. Data collection is conducted on the same target population, but respondents at one time will be different people to those in a prior year and any overlap that may occur is so rare that it cannot be considered significant. The term "trend" is used for these repeated cross-sectional surveys on different samples.

When cross-sectional surveys are repeated at regular or irregular intervals, estimates of changes can be made at the aggregate or population level. Examples include monthly labor force surveys, retail trade surveys, television and radio ratings surveys, and political opinion polls (Lavrakas, 2008). An early example of cross-sectional studies is the European Community Eurobarometer Surveys, a series of pan-European surveys undertaken for the European Commission since 1970, covering attitudes toward European integration, policies, institutions, social conditions, health, culture, the economy, citizenship, security, information technology, and the environment. A second example is the European Values Study (EVS), a large-scale, repeated cross-sectional survey research program that provides insights into the ideas, beliefs, preferences, attitudes, values, and opinions of European citizens. It has been conducted every nine years since 1981. With more than 47 participating countries, the EVS is the most comprehensive research project on human values in Europe. A third example is the European Social Survey, which was established in 2001. Every two years, face-to-face interviews are conducted with newly selected, cross-sectional samples. The survey measures attitudes, beliefs, and behavior patterns of diverse populations in more than 30 nations to understand how Europe's social, political, and moral fabric is changing.

Prospective Longitudinal Studies

Prospective longitudinal studies typically follow participants into the future; sample members are interviewed at discrete time points (e.g., every year or every few years). Surveys following persons over time can be of two broad types – cohort or panel surveys.

Cohort Studies

Cohort studies are studies in which a cohort – a group of individuals sharing some characteristic – are traced over time. A cohort has been defined as "the aggregate of individuals who experienced the same life event within the same time interval" (Ryder, 1965, p. 845). This includes birth, marriage, moment of entry in the labor market, moment of diagnosis of a particular disease, etc. In a cohort study, respondents are followed from an identical point in their life onward, generally at infrequent intervals (Dale & Davies, 1994; Laurie, 2013; Taris, 2000). One particularly important type of cohort is the "birth cohort" – the set of people who were born in the same year. Cohort studies often begin at birth but may also begin at a much later age. Thus, cohort studies gather information about a specific segment of the population, while panel studies, as we will shortly see, aim to represent the entire population.

Examples of cohort studies are the European Longitudinal Study of Pregnancy and Childhood (ELSPAC), initiated by the World Health Organization Regional Office for Europe in 1985, to identify factors influencing children's health in European countries (World Health Organization, 1999), and the UK Millennium Cohort Study (MCS), which is following the lives of around 19,000 young people born across England, Scotland, Wales, and Northern Ireland between September 2000 and January 2002. The study began with an original sample of 18,818 cohort members, and there have been seven sweeps of data collection, to date, at age 9 months and then at 3, 5, 7, 11, 17, and 22 years (2018). A further sweep of data collection is planned for age 22 years (2022). The MCS data cover topics such as parenting practices, childcare arrangements, parents' employment and education, income and poverty, family formation and dissolution, cognitive development, behavior and physical growth, and health (Connelly & Platt, 2014).

Panel Surveys

Panel surveys provide longitudinal data on a group of people, households, employers, or other social unit (termed "panel") and collect data at relatively frequent intervals (waves) depending on the design requirements. Some run over many years while others are short term, such as short panels conducted around elections to analyze individual changes of political attitudes and political behavior over the course of the campaign (Laurie, 2013). Unlike cohort studies, panel surveys commonly sample from the entire age range and collect repeated measures throughout people's life courses.

Some of the most complex panel studies are household panel surveys, which trace individuals and their households over time by gathering information about them at regular intervals. These studies collect data both from individual people (as tends to happen with cohort studies) and the whole household at each wave. Moreover, household panel studies involve both a random sample of households and all those members and subsequent co-residents, partners, and descendants who are repeatedly re-interviewed (CLOSER, 2021). However, the "unit of analysis" in virtually all longitudinal surveys is an individual person, not the family or household (Buck et al., 1995, p. 2). This is because the concept of an "individual" is stable in a longitudinal context, while families and households are dynamic because they constantly change over time.

Panel studies have been used extensively to monitor poverty and income dynamics, welfare use, social exclusion, movements into and out of the labor market, career trajectories, transitions (e.g., into/out of the labor force and from youth to adulthood), household formation and dissolution, and household change. As mentioned above, the first household panel in history was launched in the USA in 1968; the legendary PSID conducted by University of Michigan, which provided the inspiration for all subsequent household panel studies. One of the motivations for the PSID project was the assumption that poverty was self-perpetuating. The panel design offered a way to determine whether such views corresponded with reality (Elder, 1985). The results obtained encouraged a radical change in the way the phenomenon of poverty was perceived. Contrary to prevailing beliefs at the time, only a very small fraction of sample members who actually experienced poverty did so beyond a year or more. The same was true for welfare dependency – welfare recipients remained on the welfare rolls for relatively short periods of time (Coe et al., 1982; Duncan et al., 1984; Pfeffer et al., 2020; Smeeding et al., 2018).

In Europe, prospective longitudinal studies started to be implemented in the 1980s. In Germany, the first wave of the SOEP went into the field in 1984 with a sample of 5,921 households and 12,245 individuals. After the fall of the Berlin wall, the study was extended to include the former East Germany to study life-courses that had been affected by marked historical and social discontinuity; a new sample (2,179 households) was added to the original one in 1990, as well as other enlargement samples, such as migrant samples (Goebel et al., 2019). The SOEP data cover a wide range of subjects including household composition, physical and mental health, occupational and family biographies, childcare and education, employment and professional mobility, earnings, social participation and time allocation, and personal satisfaction. The SOEP is located at the German Institute for Economic Research (the Deutsches Institut für Wirtschaftsforschung [DIW]) in Berlin.

The BHPS started in 1991. The first wave consisted of 5,500 households and 10,300 individuals drawn from 250 areas of Great Britain. The sample was designed to be representative of the population (excluding Northern Ireland and North of the Caledonian Canal). Additional samples of 1,500 households in each of Scotland and Wales were added to the main sample in 1999, and in 2001 a sample of 2,000 households was added in Northern Ireland, making the panel suitable for UK-wide research. The BHPS ran from 1991 to 2009 and was extended in 2008 with the "Understanding Society" study (Buck & McFall, 2012; Platt et al., 2020). As part of wave 18, BHPS participants were asked if they would consider joining the new, larger, and more wide-ranging survey. Understanding Society is today the largest household panel survey in the world, with about 40,000 households and 50,994 individuals followed yearly since 2009. The study is based at the University of Essex Institute for Social and Economic Research.

Longitudinal-experimental studies are worth citing here. These are studies in which an initial experimental intervention is then followed up over time. As explained by Farrington et al. (2009), experiments are usually designed to investigate only immediate or short-term causal effects. However, some interventions may have long-term rather than short-term effects, and in some cases the long-term effects may differ from the short-term ones. To monitor the development of the cause–effect relationship, follow-up measurements at several different time intervals are desirable.

Prospective longitudinal data can also be drawn from surveys, official statistics, or other sources (Andreß, 2017), without personal interviews. Data can be obtained by linking together personal records from existing temporally separate data sources (e.g., administrative records gathered for official purposes or surveys such as national censuses; Buck et al., 1995). One interesting example of linked longitudinal data with administrative records is the one provided by the IAB–SOEP Migration Sample – a household survey conducted jointly by the Institute for Employment Research (IAB) in Nuremberg and the at DIW Berlin. The first survey was carried out between May and November 2013; around 2,700 households were surveyed, each containing at least one person who had migrated to Germany since 1994 or whose parents had. The sample was drawn from the Integrated Employment Biographies (IEB) sample, a database containing the entire labor market history of individuals in Germany from 1975 onward. For a subsample of the IAB–SOEP Migration Sample – only upon explicit consent of the respondents – individual data are linked to register data from the IEB. The project is aimed at overcoming limitations of previous data sets regarding the changing socio-economic structure of migration to Germany (Brücker et al., 2014).

Retrospective Surveys

In retrospective surveys, respondents are typically interviewed only once, and they are asked to remember events and circumstances of their own life course (Buck et al., 1995). Longitudinal retrospective studies can be carried out through interviews, in which participants are asked to recall personal events, or using administrative data to fill in information on past circumstances (CLOSER, 2021). Examples of retrospective questions are the following: "Since March 2020 last year, in all, how many days have you spent in a hospital or clinic as an in-patient?" or "How often have you changed your job during the last 5 years?" One key example is that of the German Life History Study (GLHS). The ten surveys of the GLHS were carried out between 1981 and 2005 by personal interviews or computer-assisted telephone interviews and collected quantitative life histories – in the form of multiple life domain event histories – from more than 12,000 respondents from eight (single or three-year) birth cohorts in West Germany born between 1919 and 1971 and five birth cohorts born between 1929 and 1971 in East Germany. The study covers more than 80 years, with the oldest cohort born in 1919 and the youngest born in 1971 and observed until 2005. Most of the surveys were retrospective; although, for the East German cohorts and the 1971 East and West German survey, a panel follow-up was conducted as well (Mayer, 2015).

Retrospective studies can also be based on secondary sources (e.g., electronic health records; Dziadkowiec et al., 2020). Kaelber et al. (2016) accumulated electronic health record data on more than 1.2 million children and adolescents (3–18 years of age) stemming from 196 pediatric primary care sites from 27 states across the USA. Participants were primary care patients with three or more visits between 1999 and 2014. The study was aimed at determining the extent to which national guidelines regarding the diagnosis of pediatric hypertension are being followed in primary care practices caring for children and adolescents.

Mixed Designs

Longitudinal research is rarely based on one method alone, but rather based on a mix of methods. Some examples of longitudinal mixed designs are the following:

- Repeated cross-sectional studies: One part of these studies are done in the form of panel studies. For example, the British Social Attitudes Survey or the Bank of Italy Survey of Household Income and Wealth are repeated regularly on a largely different sample but with a small part as a panel study. Another well-known example is the European Union Statistics on Income and Living Condition (EU-SILC) survey, implemented to study poverty and social inclusion within the EU, which provides two types of data: (1) cross-sectional data on income, poverty, social exclusion, and other living conditions; (2) longitudinal data concerning individual-level changes over time, observed periodically over a four-year period (i.e., a rotating panel design).
- Prospective studies: These studies gather information systematically using calendars and/or batteries of questions that aim to retrospectively investigate the life of the interviewee but not necessarily enquire about the same subject each time. Many panel surveys collect event history data by asking respondents retrospective questions regarding status changes, such as transitions and events that occurred in the time since the last interview (Brüderl et al., 2017). As such, event history data actually provides information on the occurrence of events (of what type, when, in what sequence) within a life course. One key example is the SHARELIFE survey, part of the Survey of Health, Ageing and Retirement in Europe (SHARE). SHARE is the largest pan-European social science panel study providing internationally comparable longitudinal micro data that allow insights in the fields of public health and socio-economic living conditions of Europeans. From 2004 until today, 480,000 in-depth interviews with 140,000 people aged 50 or older from 28

European countries and Israel have been conducted. Wave 3 (SHARELIFE) was conducted as a retrospective survey to collect information about respondents' life histories (Schröder, 2011). In SHARELIFE, retrospective data with respect to childhood, partners, children, accommodation, employment, socio-economic, and health conditions, were collected with the help of a "life history calendar." Almost 30,000 men and women across 13 European countries took part in this round of the survey.

- Cohort studies: These studies can be prospective, retrospective, or can have both a retrospective and a prospective component (ambidirectional). One good example of this is the National Child Development Study (NCDS), a birth cohort study following the lives of an initial 17,415 people born in England, Scotland, and Wales in a single week of 1958 (Power & Elliott, 2006). Since the first birth sweep, the NCDS cohort members have been followed up ten times. Data have been collected from several different sources (the midwife present at birth, parents of the cohort members, teachers, doctors, and the participants themselves) and in a variety of ways, including via paper and electronic questionnaires, clinical records, medical examinations, physical measurements, tests of ability, and educational assessments.

Qualitative Longitudinal Research

Data used in longitudinal studies may be quantitative and/or qualitative. Social change and processes can also be examined through qualitative longitudinal research (QLR). QLR involves repeated interviews conducted with the same participants over a significant period to capture temporal changes in beliefs, attitudes, and experiences at different points of their life courses (Morrow & Crivello, 2015) as well as critical moments of change and transitions (Elder & Giele, 2009). In QLR, the same people are interviewed several times in roughly fixed intervals (e.g., every two years) or around certain events (e.g., before and after childbirth; Farrall et al., 2016; Vogl et al., 2018; Winiarska, 2017). Even if QLR is rooted in a long-established tradition of qualitative temporal research – spanning the fields of social anthropology, sociological studies, and biographical research (Neale, 2019; Thomson & McLeod, 2015) – it has only recently started to systematically develop. It is from the beginning of the 2000s that studies have increasingly emerged that employ qualitative techniques in the collection and analysis of data from subjects followed over time (Hermanowicz, 2013; Thomson & Holland, 2003). There is a specific interest in using QLR to focus on children and young people, as their well-being is crucial to shape the world of tomorrow (Busse & Backeberg, 2015).

One useful example is "Inventing Adulthoods," a qualitative longitudinal study in which 100 young people, from five socially and economically contrasting areas of England and Northern Ireland, have been followed over a five-year period. The biographical material, generated in up to seven interviews with each participant, provides a unique insight into most aspects of growing up during a period of rapid social change (between 1996 and 2006 in England and between 1996 and 2010 in Northern Ireland; Henderson et al., 2006). One further example is "Italian Lives-ITA.LI," a longitudinal quantitative and qualitative research project on Italian families carried out by the Department of Sociology and Social Research of the University of Milan-Bicocca. Its aim is to monitor social change in Italy, offering high-quality data to researchers. The quantitative study began in 2019 and involves all household members aged 16 and over living in approximately 4,900 families, selected from more than 278 Italian municipalities, using a probabilistic sampling method. During the first wave, individual life courses were reconstructed through retrospective questions. The qualitative study aims to collect data for analyzing the everyday experiences of young people. This includes their work experiences, friendships and intimate relationships, intergenerational relations, housing issues, time use and leisure activities, and young people's agency. The survey involves a group of women and men aged between 23 and 29 and is carried out in several waves of interviews held at regular intervals. The participants are extracted from the quantitative survey sample and are selected for interview on a voluntary basis.

Longitudinal Research: Benefits and Challenges

The debate on advantages and disadvantages of longitudinal research dates back to the 1920s due to the criticisms of cross-sectional methods failing to properly explain growth (Rajulton, 2001). More elaborate discussion had to wait for four more decades, beginning in the 1960s (Rajulton & Ravanera, 2000). Today, it is widely recognized that longitudinal research has clear pros and cons (Caruana et al., 2015; Lynn, 2009). Even though dynamic data offer a highly innovative tool for the analysis of social phenomena, they do, nonetheless, have certain inherent disadvantages that the researcher should keep in mind. Below we summarize the main advantages and disadvantages of each longitudinal design. As pointed out by some scholars (e.g., Buck et al., 1995), choosing the most appropriate survey design requires assessment of the benefits of the different sorts of information provided and the different costs required to derive them.

Repeated Cross-Sectional Design

As mentioned above, a cross-sectional study analyzes a cross-section of the population at a specific point in time.

Strengths

Cross-sectional studies have several benefits. Their one-off nature makes such studies easier to organize, relatively cheap, and less time-consuming than other types of research. They also have the advantage of immediacy, allowing researchers to collect a great deal of information quickly and offering instant results. Repeated cross-sectional surveys are suitable for measuring prevalence and change over time at the population level (McManus, 2020) and are used to study trends. According to Hagenaars (1990, p. 271), trend studies have some advantages over panel and cohort studies as trend data are more readily available and can be analyzed in a simpler way than cohort and panel data. The investigation of long-term social change, in particular, has to rely on trend rather than panel data. Moreover, while cross-sectional studies cannot be used to determine causal relationships, they can provide a useful starting point to further research. If representative samples are present in consecutive years of a survey, it is possible to compare changes in the behavior or circumstances of different groups (e.g., a comparison between the incidence of poverty and the characteristics of the population below the poverty line at time *t* and at time *t* + 1 or between the pool of employed and unemployed in two different years).

Limitations and Challenges

However, because questions are asked of a new sample every time, these studies only offer a means for analyzing changes for population groups (also known as aggregate change – the net effect of all the changes; Firebaugh, 1997; Rafferty et al., 2015). They cannot be used to look at individual change, shedding little light on who has changed, how, or why. Social and behavioral scientists should be very careful when attempting to extrapolate longitudinal inferences on the basis of analyses of cross-sectional data as they have to, implicitly, assume that the process being studied is in some sort of equilibrium. Consequently, it should come as no surprise that conclusions drawn based on cross-sectional data have often been challenged by analyses based on longitudinal data (Davies, 1994; Ghellini & Trivellato, 1996).

Panel Design

Strengths

Panel data offer clear advantages for the study of social dynamics and the connection between individual, family, and social change. While cross-sectional studies do not reveal whether any changes that show up should be attributed to new individuals or to a real change in behavior, panel studies resolve this problem because they, periodically, gather information about the same subjects. Because sample members are surveyed at successive time points, it is possible to investigate how individual outcomes are related to earlier circumstances. Prospective studies help to unravel the nature of change at an individual level and are thus considered preferable when analyzing microsocial change (Dale & Davies, 1994; Janson, 1990; Longhi & Nandi, 2015; Magnusson et al., 1991; Rose, 2000). The longitudinal structure of the data makes it also possible to contextualize changes within the institutional, cultural, and social environments that surround the individual and shape the course of his or her life (Ruspini, 2002, 2008). For example, some scholars (Kühne et al., 2020) have argued that household panel data seem ideally suited for research on the short-term and long-term effects of extreme events on individuals and households as well as to understand how their micro-level consequences translate into complex social phenomena and macro-level structural change Panel surveys also provide the most reliable data on changes in beliefs, values, and attitudes because longitudinal measures are collected while the subjective states actually exist.

Limitations and Challenges

Prospective longitudinal studies have certain well-known disadvantages (Blossfeld & Rohwer, 2013; Magnusson & Bergmann, 1990). The collection of panel data is much more costly than the collection of cross-sectional data. They are very expensive both in terms of the money and of the time and energy they require; for this reason, they are usually carried out by large research organizations and often need governmental support. The higher costs are derived from the fact that researchers must follow the subjects over time. They must track people who form a new family, who move houses, or move to another municipality so that events, such as births, divorces, children leaving home, new marriages, and cohabitations will be reflected in the sample in the same proportion as they are to be found in the general population. The rules that decide which household members are still surveyed after they leave the household or which respondents are still surveyed when households split up are called "following rules," and they must be decided upon at the design stage. Typically, if a respondent who is followed forms a new household, all new household members are interviewed while living with that member of the sample. However, respondents who move into institutions (e.g., elderly care homes or prisons) are generally not interviewed. The Australian HILDA panel is an exception insofar as it follows respondents moving into nursing homes and other non-private dwellings (but not into prisons; Schonlau et al., 2011). Moreover, there is also a need to preserve the research team over the duration of the study (van der Kamp & Bijleveld, 1998). Finally, it is important to remember that time must elapse before any analysis of social change can result, and long-term in-depth analyses of individual and social processes require data gathered from a considerable number of waves. To study change, one should collect at least three waves of data, and a multiple-wave study is better (Ployhart & Ward, 2011).

Panel data also suffer from attrition problems. Attrition occurs when respondents leave the panel after having participated in one or more consecutive waves and results in a diminishing number of study respondents. Attrition occurs for various reasons, including a refusal to continue, physical incapacity of the respondent to provide information, death or emigration, and/or failure to follow up sample cases (Lepkowski & Couper, 2002). This thinning process is not random – some individuals are more likely to drop out of a study than others. There are well-established risk factors – such as lower education, low income, declining health, old age – for attrition of study participants.

If not controlled, selective attrition can negatively affect the representativeness of the sample, misleading estimates of change measures and distorting conclusions drawn based on information supplied by that section of the sample that remains. Indeed, the capability of panel surveys to capture change depends on the extent to which the sample remains representative of the study population over time (Fumagalli et al., 2012). The best way to counter the problem of attrition during the period of observation planned is to ensure that it starts with a high-quality initial sample (Duncan, 2000). There is also a need to plan and adopt specific techniques to successfully follow sample members over time and maintain a high level of participation between one wave and the next; typically, longitudinal studies lose most of their sample between the first and second rounds of data collection. Different methods can reduce attrition in longitudinal surveys (Andreß, 2017). These include tracking procedures to keep in touch with respondents over the course of the study, imputing and re-weighting to reduce unit non-response bias, drawing refreshment samples, or implementing a so-called "rotating panel" design – equally sized sets of sample units are added to the sample at each successive wave to correct distortions that may have arisen within the sample between time *t* and time *t* + 1 (e.g., one-sixth of the sample retire and are replaced by an equal number of employed). For example, the SOEP has added several refreshment samples over the years.

There is also higher risk of error than in cross-sectional data because errors accumulate over time (Fuller, 1987). For example, if data about income gathered at time *t* have errors, this could lead to false transitions appearing concerning phenomena such as poverty or unemployment. Panel studies also tend to influence the phenomena that they are hoping to observe; "panel conditioning"is another effect sometimes observed in repeated surveys. Repeated questioning of panel members can influence their survey responses, either by altering the behavior reported or by changing the quality of the responses given. The reason may be that respondents have acquired new information in the meantime or that they have had new experiences during the time that has elapsed between one wave and the next (Duncan, 2000). Moreover, a particular situation or event that occurred at the time when the information is collected may also distort individual answers.

Another issue not to be forgotten is that panel data offer information that is related only to predetermined points in time. That is, data are usually gathered annually (i.e., at discrete time points). Thus, the researcher cannot know about the course and evolution of events in the period that has elapsed between one collection time and the next. Furthermore, prospective studies are often limited to a few waves only and, consequently, cover only a short period of time. One example of this is the European Community Household Panel (ECHP), a longitudinal household survey covering 14 EU member states (Belgium, Denmark, Germany, Ireland, Greece, Spain, France, Italy, Luxembourg, the Netherlands, Austria, Portugal, Sweden, and the United Kingdom). After a total duration of eight years (1994–2001), Eurostat decided to stop the ECHP project and to replace it in 2003 with a new instrument, the already mentioned EU-SILC survey.

There are also problems that are inherent in the structure of the panels themselves. First, panel data files are usually extremely large as they accumulate a large amount of data over the years. Most existing household panels have initial samples of around 5,000 households and more than 10,000 individuals. The high level of complexity of the structure of household panel studies is also a problem. These studies are complex in the sense that they consist of several different data files with differing focuses – some referring to the particular households studied at particular waves, some referring to individuals, some referring to particular events that the interviewees have experienced in successive years or waves. For example, the interview methodology of the SOEP is based on a set of questionnaires for households and individuals aged 16 and over. The SOEP questionnaires are designed so that people in a SOEP household can be analyzed from birth to adulthood and throughout the rest of their lives. A rather stable set of core questions is asked every year, enhanced by topical modules and rotating modules on topics such as wealth, neighborhood, family and social networks, social security, and time use. Additionally, one person (household reference person) is asked to answer a household-related questionnaire covering information on housing, housing costs, and different sources of income (e.g., social assistance or housing allowances). This questionnaire also includes questions on children up to the age of 16 living in the household, mainly concerning day care, kindergarten, and school attendance.

In other words, the structure of household panel data makes it possible to combine two separate units of analysis (family and individual) and to create longitudinal files (by linking one wave to another using unique individual and household identifiers) on the basis of either prospective or retrospective longitudinal information that has been gathered at either the aggregate or the individual level. The user documentation is thus crucial to making longitudinal analysis both easier and more straightforward. It should contain essential information required for the analysis of data and information, as well as information that will assist users when linking and aggregating data across waves (Freed Taylor, 2000).

Finally, the analysis of panel data is, in itself, highly complex and needs specific statistical procedures (Caruana et al., 2015; Devaney & Rooney, 2018; Rajulton, 2001). One last consideration, as mentioned above, with the explicit consent of survey respondents, longitudinal data can be linked to administrative data such as hospital episodes, benefits, or educational records. The linkage makes it also possible to obtain precise information on wages and salaries, employment, unemployment and benefit receipt, as well as many other variables that are particularly relevant to labor market issues (Brücker et al., 2014). However, there are also significant problems with using record linkages. First, linkage may simply be impossible, as a result of confidentiality or privacy restrictions relating to collection of the original data. A second problem is that analysis is constrained by the coverage of the variables contained in the original surveys, often rather limited (e.g., tax records) (Buck et al., 1995).

Retrospective Design

Strengths

Retrospective designs are faster to conduct and less expensive than prospective longitudinal data, as they are usually gathered during one single wave. The advantages of this method are its simplicity, cost (i.e., there is only a single interview and respondents do not have to be tracked), and the immediate availability of longitudinal information – the researcher does not have to wait for a second interview to detect change (Buck et al., 1995). Retrospective data can also be very rich because respondents are asked to remember events and aspects of their own life-courses.

Typically, this is done for one time followed by another, beginning with the current situation and taking respondents backward in time.

Limitations and Challenges

Retrospective surveys, however, have clear limitations, both due to the necessarily simplified form in which they are forced to reconstruct life experiences and because of memory biases when trying to recall past events (Blossfeld & Rohwer, 2013; Dex, 1995; Hakim, 1987; Taris, 2000). Hence, retrospective surveys are usually limited to significant but infrequent life events, such as births, marriages, divorces, and job changes (Rose, 2000). In general, the quality of the data diminishes the further back in time the interviewee is asked to go; the longer the recall period is, the more unreliable retrospective data tend to be.

Another disadvantage is linked to the quantity of information that an individual can remember on one occasion (i.e., when the retrospective interview is carried out). Many participants simply forget things about events, feelings, or considerations; even when an event has not been wholly forgotten, they may have trouble recalling it (due to memory loss and retrieval problems). One particular type of memory error occurs when respondents omit relevant pieces of information. Respondents may be unable to recall a particular item, or they may be unable to distinguish one item from another in their memories (Linton, 1982). Even if all relevant events have been correctly remembered, if asked when they happened, respondents tend to report events as having taken place more recently than they actually did (forward telescoping; Ziniel, 2008). The inverse may also occur; some participants place events further away in the past than they actually happened (backward telescoping, Ziniel, 2008). Thus, only a period that has a well-defined limit, usually the preceding wave, should be used. This helps to reduce the effects of telescoping and, to some extent, to keep a check on them (Janson, 1990; Sudman & Bradburn, 1982).

Retrospective questions concerning cognitive and affective states and attitudes are particularly problematic; it can be very difficult for interviewees to accurately remember the changes related to particular states of mind, how long these states lasted, and the precise order in which they took place. Finally, in some other areas – such as income or state of health – it is quite difficult to collect accurate information retrospectively due to the fallibility of memory (e.g., information about monthly earnings, blood pressure, weight loss or gain, etc.).

Suggestions and Conclusions

Longitudinal data, either prospective or retrospective, offer several advantages. They make it possible to analyze the duration of social phenomena and highlight differences or changes, between one period and another, in the values of one or more variables. They can be used to examine the flows into and out of a situation, such as poverty, illness, or unemployment (Duncan & Kalton, 1987; Rose, 1993, 2000). They also provide information about the ordering of events in time, allowing antecedents to be specified and consequences identified, as needed to draw conclusions about causes (Leisering & Walker, 1998). Longitudinal data allow the identification of sleeper effects (i.e., connections between events and transitions that are widely separated in time because they took place in very different periods) in the relation between childhood, adulthood, and old age (Caruana et al., 2015; Elder, 1985; Hakim, 1987). For example, the positive or negative experience of old age has much to do with experiences and resources accumulated throughout life. Longitudinal research has also suggested that as children of divorce enter adulthood, they may be more likely than other subjects to have difficulties in relationship formation and maintenance and have fears regarding betrayal and abandonment in couple relationships (Sarigiani & Spierling, 2011; Wallerstein et al., 2000).

However, conducting longitudinal research can be a demanding task, and numerous variables are to be considered, and adequately controlled, when embarking on such a project – particularly in view of the protracted nature of such a commitment (Caruana, 2005). Longitudinal studies, especially panel studies, need an appropriate infrastructure for the actual duration of the study to withstand the test of time. The organizational costs of longitudinal research are tremendous; not only must it be ensured that the same subjects can be traced repeatedly over their life-course, but the research team must be kept constant over the duration of the study (van der Kamp & Bijleveld, 1998). It is also crucial to carefully preserve all data and all related documents to keep track of the data production process and its possible changes. A further important aspect is to control and minimize attrition. The best way to counter this problem is to ensure that the study has a high-quality initial sample, has clear rules to follow up the sample over time and to update the original sample, and adopts effective strategies to maintain a high level of participation/response (Andreß, 2017; Duncan, 2000; Ghellini & Trivellato, 1996; Rose, 2000; Ruspini, 2002). The fundamental rule to set when defining a reference population or populations, longitudinally, is to follow up all the original members of the sample and all those born to these original members.

As regards the choice of longitudinal designs, what guidelines can be offered to researchers? If there is no interest in causal relationships or if causal and temporal order are known, then cross-sectional data may be enough (van der Kamp & Bijleveld, 1998). However, repeated cross-sectional designs may be appropriate if it is thought that the problem of panel conditioning may arise as a result of repeated interviewing or observation. On the other hand, if a study aims to discover causal mechanisms, longitudinal panel data provides a stronger foundation for causal inferences. Panel studies contain measures of variables for each unit at different time points. Hence, it is possible to use information about prior as well as current values of variables in constructing and estimating causal models (Finkel, 1995).

If change is to be measured over a long time span, then a prospective panel is the most appropriate design for the study. If change is to be measured only over a relatively short time (weeks or months), a retrospective design may be appropriate for data concerning events or behavior, but probably not for attitudes or beliefs. Finally, to combine the strengths of panel designs and the virtues of retrospective studies, a mixed design employing a follow-up and a follow-back strategy seems appropriate.

References

Andreß, H-J. (2017). The need for and use of panel data. *IZA World of Labor*, *352*. https://www.doi.org/10.15185/izawol.352

Blossfeld, H. P. & Rohwer, G. (2013). *Techniques of Event History Modeling. New Approaches to Causal Analysis*. Routledge.

Brücker, H., Kroh, M., Bartsch, S., Goebel, J., et al. (2014). *The New IAB–SOEP Migration Sample: An Introduction into the Methodology and the Contents. SOEP Survey Papers 216: Series C.* German Institute for Economic Research (DIW)/German Socio-Economic Panel (SOEP).

Brüderl, J., Castiglioni, L., Volker, L., Pforr, K., & Schmiedeberg, C. (2017). Collecting event history data with a panel survey: Combining an electronic event history calendar and dependent interviewing. *Methods, Data, Analyses*, *11*(1), 45–66.

Buck, N. & McFall, S. (2012). Understanding society: design overview. *Longitudinal and Life Course Studies*, *3*(1), 5–17.

Buck, N., Ermisch, J., & Jenkins, S. (1995). Choosing a longitudinal survey design: the issues. Paper ESRC Research Centre on Micro-Social Change, University of Essex. Available at: www.iser.essex.ac.uk/files/occasional_papers/pdf/op96-1.pdf.

Busse, B. & Backeberg, L. (2015). Longitudinal research on children and young people in Europe and beyond. In G. Pollock, J. Ozan, H. Goswami, G. Rees, & A. Stasulane (eds.), *Measuring Youth Well-Being. How a Pan-European Longitudinal Survey Can Improve Policy* (pp. 71–89). Springer.

Caruana, E. J., Roman, M., Hernández-Sánchez, J., & Solli, P. (2015). Longitudinal studies. *Journal of Thoracic Disease*, *7*(11), 537–540.

CLOSER (2021). Learning hub. Available at: https://learning.closer.ac.uk/learning-modules/introduction/types-of-longitudinal-research/panel-studies/.

Coe, R. D., Duncan, G. J., & Hill, M. S. (1982). Dynamic aspects of poverty and welfare use in the United States. Paper presented at the Conference on Problems of Poverty, Clark University, Worcester, MA, August.

Connelly, R. & Platt, L. (2014). Cohort profile: UK Millennium Cohort Study (MCS). *International Journal of Epidemiology*, *43*(6), 1719–1725.

Dale, A. & Davies, R. B. (eds.) (1994). *Analysing Social and Political Change. A Casebook of Methods*. SAGE Publications.

Davies, R. B. (1994). From cross-sectional to longitudinal analysis. In A. Dale & R. B. Davies (eds.), *Analysing Social and Political Change. A Casebook of Methods* (pp. 20–40). SAGE Publications.

Devaney, C. & Rooney, C. (2018). *The Feasibility of Conducting a Longitudinal Study on Children in Care or Children Leaving Care within the Irish Context*. UNESCO Child and Family Research Centre, National University of Ireland.

Dex, S. (1995). The reliability of recall data: a literature review. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 49(1), 58–89. https://www.doi.org/10.1177/075910639504900105

Doll, R. (2001). Cohort studies: History of the method II. Retrospective cohort studies. *Sozial und Präventivmedizin*, *46*, 152–160.

Duncan, G. J. (2000). Using panel studies to understand household behavior and well-being. In D. Rose (ed.), *Researching Social and Economic Change. The Uses of Household Panel Studies* (pp. 54–75). Routledge.

Duncan, G. J., Coe, R. D., Corcoran, M. E., et al. (1984). *Years of Poverty, Years of Plenty: The Changing Economic Fortunes of American Workers and Families*. Institute for Social Research, University of Michigan.

Dziadkowiec, O., Durbin, J., Jayaraman Muralidharan, V., Novak, M., & Cornett, B. (2020). Improving the quality and design of retrospective clinical outcome studies that utilize electronic health records. *HCA Healthcare Journal of Medicine*, *1*(3), article 4.

Elder, G. H., Jr. (1985) Perspectives on the life-course. In G. H., Jr. Elder (ed.) *Lifecourse Dynamics. Trajectories and Transitions, 1968–1980* (pp. 23–49). Cornell University Press.

Elder, G. H., Jr. & Giele, J. Z. (2009). *The Craft of Life Course Research*. The Guilford Press.

Farrall, S., Hunter, B., Sharpe, G., & Calverley A. (2016). What 'works' when retracting sample members in a qualitative longitudinal study? *International Journal of Social Research Methodology*, *19*(3), 287–300.

Farrington, D. P., Loeber, R., & Welsh, B. C. (2009). Longitudinal-experimental studies. In A. R. Piquero, & D. Weisburd (eds.), *Handbook of Quantitative Criminology* (pp. 503–518). Springer.

Finkel, S. E. (1995). *Causal Analysis with Panel Data*. SAGE Publications.

Firebaugh, G. (1997). *Analyzing Repeated Surveys*. SAGE Publications.

Freed Taylor, M. (2000). Dissemination issues for panel studies: Metadata and documentation. In D. Rose (ed.), *Researching Social and Economic Change. The Uses of Household Panel Studies* (pp. 146–162). Routledge.

Fuller, W. A. (1987). *Measurement Error Models*. John Wiley & Sons.

Fumagalli, L., Laurie, H., & Lynn, P. (2012) Experiments with methods to reduce attrition in longitudinal surveys. *Journal of the Royal Statistical Society. Series A*, *176*(2), 499–519.

Ghellini, G. & Trivellato, U. (1996) Indagini panel sul comportamento socio-economico di individui e famiglie: una selezionata rassegna di problemi ed esperienze. In C. Quintano (ed.) *Scritti di Statistica Economica 2*. Rocco Curto Editore.

Giroux, É. (2011). The origins of the prospective cohort study: American cardiovascular epidemiology and the Framingham Heart Study. *Revue d'histoire des sciences*, *2*(2), 297–318.

Goebel, J., Grabka, M., Liebig, S., et al. (2019). The German Socio-Economic Panel (SOEP). *Journal of Economics and Statistics (Jahrbuecher fuer Nationaloekonomie und Statistik)*, *239*(2), 345–360.

Hagenaars, J. A. (1990). *Categorical Longitudinal Data; Log-Linear Panel, Trend, and Cohort Analysis*. SAGE Publications.

Hakim, C. (1987). *Research Design. Strategies and Choices in the Design of Social Research*. Allen and Unwin.

Henderson, S. J., Holland, J., McGrellis, S., Sharpe, S., & Thomson, R. (2006). *Inventing Adulthoods: A Biographical Approach to Youth Transitions*. SAGE Publications.

Hermanowicz, J. C. (2013). The longitudinal qualitative interview. *Qualitative Sociology*, *36*, 189–208.

Janson, C-G. (1990). Retrospective data, undesirable behavior, and the longitudinal perspective. In D. Magnusson & L.R. Bergman (eds.), *Data Quality in Longitudinal Research* (pp. 100–121). Cambridge University Press.

Kaelber, D. C., Liu, W., Ross, M., et al. (2016). Diagnosis and medication treatment of pediatric hypertension: A retrospective cohort study. *Pediatrics*, *138*(6), e20162195. https://www.doi.org/10.1542/peds.2016-2195

Kühne, S., Kroh, M., Liebig, S., & Zinn, S. (2020). The need for household panel surveys in times of crisis: The case of SOEP-CoV. *Survey Research Methods*, *14*(2),195–203.

Laurie, H. (2003). From PAPI to CAPI: Consequences for data quality on the British Household Panel Survey. *Working Papers of the Institute for Social and Economic Research*, paper 2003–14.

Laurie, H. (2013). Panel studies. *Oxford Bibliographies in Sociology*. https://www.doi.org/10.1093/obo/9780199756384-0108

Lavrakas, P. J. (2008). *Encyclopedia of Survey Research Methods*. SAGE Publications.

Leisering, L. & Walker, R. (1998). Preface. In L. Leisering & R. Walker (eds.), *The Dynamics of Modern Society* (pp. x–xvii). The Policy Press.

Lepkowski, J. M. & Couper, M. P. (2002). Nonresponse in the second wave of longitudinal household surveys. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (eds.). *Survey Nonresponse* (pp. 259–272). John Wiley & Sons.

Linton, M. (1982). Transformations of memory in everyday life. In U. Neisser (ed.), *Memory Observed. Remembering in Natural Contexts*. W. H. Freeman.

Lynn, P. (ed.) (2009). *Methodology of Longitudinal Surveys*. John Wiley & Sons,

Longhi, S. & Nandi, A. (2015). *A Practical Guide to Using Panel Data*. SAGE Publications.

Magnusson, D. & Bergman, L. R. (eds.) (1990). *Data Quality in Longitudinal Research*. Cambridge University Press.

Magnusson, D., Bergman, L. R., Rudinger, G., & Torestad, B. (eds.) (1991). *Problems and Methods in Longitudinal Research: Stability and Change*. Cambridge University Press.

Mayer, K. U. (2015). The German life history study: An introduction. *European Sociological Review*, 31(2), 137–143.

McManus, S. (2020). Using repeated cross-sectional surveys to measure trends in rates of self-harm. *Sage Research Methods Cases: Medicine and Health*. https://www.doi.org/10.4135/9781529733679

Menard, S. (2002). *Longitudinal Research*, 2nd ed. SAGE Publications.

Morrow, V. & Crivello, G. (2015). What is the value of qualitative longitudinal research with children and young people for international development? *International Journal of Social Research Methodology*, *18*(3), 267–280.

Neale, B. (2019). *What Is Qualitative Longitudinal Research?* Bloomsbury Academic.

Nesselroade, J. R., & Baltes, P. B. (1979). *Longitudinal Research in the Study of Behavior and Development*. Academic Press.

Pfeffer, F. T., Fomby, P., & Insolera, N. (2020). The longitudinal revolution: Sociological research at the 50-year milestone of the Panel Study of Income Dynamics. *Annual Review of Sociology*, *46*(1), 83–108.

Platt, L., Knies, G., Luthra, R., Nandi, A., & Benzeval, M. (2020). Understanding society at 10 years. *European Sociological Review*, *36*(6), 976–988.

Ployhart, R. E. & Ward, A.-K. (2011). The "quick start guide" for conducting and publishing longitudinal research. *Journal of Business and Psychology*, *26*(4), 413–422.

Power, C. & Elliott, J (2006). Cohort profile: 1958 British birth cohort (National Child Development Study). *International Journal of Epidemiology*, *35*(1), 34–41.

Rafferty, A., Walthery, P., & King-Heşe, S. (2015). *Analysing Change Over Time: Repeated Cross-Sectional and Longitudinal Survey Data*. UK Data Service, University of Essex and University of Manchester.

Rajulton, F. (2001). The fundamentals of longitudinal research: An overview special issue on longitudinal methodology. *Canadian Studies in Population*, *28*(2), 169–185.

Rajulton, F. & Ravanera, Z. R. (2000). Theoretical and analytical aspects of longitudinal research. *PSC Discussion Papers Series*, *14*(5), article 1. https://ir.lib.uwo.ca/pscpapers/vol14/iss5/1.

Rose, D. (ed.). (2000). *Researching Social and Economic Change: The Uses of Household Panel Studies*. Routledge.

Ruspini, E. (2002). *Introduction to Longitudinal Research*. Routledge.

Ruspini, E. (2008). Longitudinal research. An emergent method in the social sciences. In S. N. Hesse-Biber & P. Leavy (eds.), *Handbook of Emergent Methods* (pp. 437–460). The Guilford Press.

Ryder, N. B. (1965). The cohort as a concept in the study of social change. *American Sociological Review*, *30*(6), 843–861.

Samet, J. M. & Muñoz A. (1998). Evolution of the cohort study. *Epidemiologic Reviews*, *20*(1), 1–14.

Sarigiani, P. A. & Spierling, T. (2011). Sleeper effect of divorce. In Goldstein S. & Naglieri J. A. (eds.), *Encyclopedia of Child Behavior and Development*. Springer. https://doi.org/10.1007/978-0-387-79061-9_2666

Schonlau, M., Watson, N., & Kroh, M. (2011). Household survey panels: How much do following rules affect sample size? *Survey Research Methods*, *5*(2), 53–61.

Schröder, M. (2011). *Retrospective Data Collection in the Survey of Health, Ageing and Retirement in Europe. SHARELIFE Methodology*. Mannheim Research Institute for the Economics of Aging (MEA).

Smeeding, T. M. (2018). The PSID in research and policy. *Annals of the American Academy of Political and Social Science*, *680*(1), 29–47. https://doi.org/10.1177/0002716218798802

Sontag, L. (1971). The history of longitudinal research: Implications for the future. *Child Development*, *42*(4), 987–1002.

Sudman, S. & Bradburn, N.A. (1982). *Asking Questions*. Jossey-Bass Publishers.

Taris, T. W. (2000). *A Primer in Longitudinal Data Analysis*. SAGE Publications.

Terman, L. M. et al. (1925). *Genetic Studies of Genius. Volume I. Mental and Physical Traits of a Thousand Gifted Children*. Stanford University Press.

Terman, L. M. et al. (1929). *Genetic Studies of Genius. Volume II. [Authored by C. M. Cox] The Early Mental Traits of Three Hundred Geniuses*. Stanford University Press.

Terman, L. M. et al. (1930). *Genetic Studies of Genius. Volume III. [Authored by B. S. Burks, D. W. Jensen, & L. M. Terman] The Promise of Youth: Follow-Up Studies of a Thousand Gifted Children*. Stanford University Press.

Thomson, R. & Holland, J. (2003). Hindsight, foresight and insight: The challenges of longitudinal qualitative research. *International Journal of Social Research Methodology*, *6*(2), 233–244.

Thomson, R. & McLeod, J. (2015). New frontiers in qualitative longitudinal research: An agenda for research. *International Journal of Social Research Methodology*, *18*(3), 243–250.

Tsao, C. W. & Vasan, R. S. (2015). Cohort profile: The Framingham Heart Study (FHS): overview of milestones in cardiovascular epidemiology. *International Journal of Epidemiology*, *44*(6), 1800–1813.

van der Kamp L. J. T. & Bijleveld C. C. J. H. (1998). Methodological issues in longitudinal research. In C. C. J. H. Bijleveld & L. J. Th. van der Kamp (eds.), *Longitudinal Data Analysis. Designs, Models and Methods* (pp. 1–45). SAGE Publications.

Venkatesh, A. & Vitalari, N. (1991). Longitudinal surveys in information systems research: An examination of issues, methods, and applications. In K. L. Kraemer (ed.), *The Information Systems Research Challenge: Survey Research Methods* (pp. 115–144). Harvard Business School Press.

Voelkle, M. C. & Adolf, J. (2015). History of longitudinal statistical analyses. In N. Pachana (ed.), *Encyclopedia of Geropsychology*. Springer. https://doi.org/10.1007/978-981-287-080-3_135-1

Vogl, S., Zartler, U., Schmidt, E-M., & Rieder, I. (2018). Developing an analytical framework for multiple perspective. Qualitative longitudinal interviews (MPQLI). *International Journal of Social Research Methodology*, *21*(2), 177–190.

Wall, W. D. & Williams, H. L. (1970). *Longitudinal Studies and the Social Sciences*. Heinemann.

Wallerstein, J. S., Lewis, J. M., & Blakeslee, S. (2000). *The Unexpected Legacy of Divorce: A 25 Year Landmark Study*. Hyperion.

Winiarska, A. (2017). Qualitative longitudinal research: Application, potentials and challenges in the context of migration research, Centre of Migration Research (CMR) Working Paper 103/161, Warsaw: University of Warsaw. Available at: www.econstor.eu/bitstream/10419/180968/1/1018535470.pdf.

World Health Organization (1999). European longitudinal study of pregnancy and childhood (ELSPAC): Report on a WHO meeting, Bristol, 13–18 September 1999. WHO Regional Office for Europe. Available at: https://apps.who.int/iris/handle/10665/108279.

Ziniel, S. (2008). Telescoping. In P. J. Lavrakas (ed.), *Encyclopedia of Survey Research Methods*,: SAGE Publications. Available at: https://methods.sagepub.com/reference/encyclopedia-of-survey-research-methods/n579.xml?fromsearch=true.

18 Online Research Methods

Kevin B. Wright

Abstract

This chapter examines four prominent online research methods – online surveys, online experiments, online content analysis, and qualitative approaches – and a number of issues/best practices related to them that have been identified by scholars across a number of disciplines. In addition, several platforms for conducting online research, including online survey and experimental design platforms, online content capture programs, and related quantitative and qualitative data analysis tools, are identified in the chapter. Various advantages (e.g., time saving, cost, etc.) and disadvantages (e.g., sampling issues, validity and privacy issues, ethical issues) of each method are then discussed along with best practices for using them when conducting online research.

Keywords: Online Surveys, Online Experimental Designs, Online Content Analysis, Online Interviews and Focus Groups, Online Data Capture, Online Data Analysis

Introduction

The way people communicate is continuously changing in the digital age, and researchers need to adapt when studying human behavior. Scholars from a variety of academic disciplines have embraced and benefited from switching from traditional face-to-face research methods to online methods, including online survey research, online experiments, online content analysis, and online qualitative approaches (e.g., Hall et al., 2020; Pechey & Marteau, 2018; Wang et al., 2015; Wright et al., 2019. With the advent of Internet-based research, many scholars were initially skeptical about the efficacy of conducting research online. However, research stemming from web-based methods is now increasingly common and being published in major disciplinary journals (Babbie, 2020; Skitka & Sargis 2006; Wright, 2005, 2017). The development of online survey/experiment platforms (e.g., Qualtrics and Gorilla) have allowed researchers to create online surveys and experimental designs that can transcend traditional data collection restrictions (e.g., mail surveys and university laboratory settings) and ease access to relatively diverse (and sometimes nationally representative) participant pools (Simmons & Bobo, 2015; Weinberg et al., 2014).

In addition, researchers are using other platforms to develop new paradigms for research, including the use of online "big data" collection and analysis. Big data web capture and analysis programs (e.g., Python), which allow for the analysis of

https://doi.org/10.1017/9781009010054.019 Published online by Cambridge University Press

large-scale, rapidly generated (often in real time) data from web content (Gandomi & Haider, 2015), offer researchers several advantages compared to traditional (offline) research methods. Online content analysis has benefitted from innovations like autocoding software (e.g., NVivo, Atlas), website and social media analytics programs (e.g., Radian6), and online social network analysis (NodeXL). For example, integrating content analysis with social network analysis programs can help researchers observe natural participant communication and message dissemination patterns across social media platforms (e.g., Twitter, Facebook, etc.) while simultaneously accounting for important demographic and social media use variables. Such approaches have led to new frontiers of online research.

The diversity and range of topics that have been studied in online settings, over the past 20 years, have expanded considerably. Researchers have found innovative ways to use new technologies/media to help them better understand many facets of human behavior in online social settings. For example, scholars from many disciplines have developed online studies of phenomena, as diverse as online marketing and multiteam systems (Mason & Watts, 2012; Shen et al., 2016), online support communities for numerous health issues (Rains et al., 2015; Wagg et al., 2019; Wright, 2016), and “citizen science” websites, in an effort to collect and analyze large-scale data (Aristeidou et al., 2017; Armstrong et al., 2020).

This recent online research activity is impressive given the fact that online research was in its nascent stages a relatively short time ago. The first online surveys began to appear in the 1990s, and they were somewhat cumbersome for researchers and participants (Wright, 2005). Early online surveys typically made use of web forms (i.e., plain HTML forms) with questions arranged one after another on a single web page. In the 2000s, the introduction of Web 2.0 paved the way for the development of more sophisticated platforms, software, and services that allowed researchers to better study online populations. For example, platforms like SurveyMonkey and Qualtrics began including interactive features in online surveys and experiments (e.g., feedback conditioned by responses on the same web page), which can be enabled on the participant side (i.e., on the respondent’s device) with the use of special browser scripts. In the last decade, small portable devices, such as smartphones and tablets, have been increasingly used by consumers to participate in web surveys and online experiments. This has generated new issues in the design of web questionnaires, such as how to manage smaller screen sizes for presenting online survey or experimental content (as compared to a desktop computer).

This chapter explores a number of online research methods and issues related to them that have been identified by scholars across a number of disciplines. Toward that end, the chapter examines these advantages and disadvantages more broadly prior to discussing the four common online research methods – online surveys, online experiments, online content analysis, and qualitative approaches to studying online populations. Many of these considerations stem from the author’s experience as an online researcher for over 20 years. However, given the rapid growth of online research methods, a full consideration of the many types of online research methods and related issues is beyond the scope of this chapter.

Advantages of Online Research Methods

Cost

In the early days of online survey research, researchers quickly learned that online questionnaires do not have to be printed and sent by mail (Wright, 2005). Moreover, online surveys allow interviewers to circumvent travel needed to recruit participants in person or to hire assistants to reach them via phone. Platforms like Qualtrics make it easy to export statistics from online surveys and experiments into statistical analysis programs (e.g., SPSS or SAS). Online surveys are typically much cheaper than more conventional survey data collection methods when you factor in the cost of materials, recruiting and training interviewers, and personnel for traditional data entry (Wright, 2005). Online experiments allow researchers to move beyond a physical laboratory when recruiting participants (who may be willing to participate in an online experiment more cheaply than if they had to come to a university building or another laboratory setting). Social media platforms provide an opportunity for researchers to analyze a vast amount of online content as well as observe naturally occurring online conversations on various platforms.

Time and Ease

The next appealing advantage of online research methods is the ease and speed of the data collection process (Wright, 2005).This is especially important for time-sensitive studies (e.g., political studies during election campaigns). Computational technology has improved the effectiveness and efficiency of methods for collecting and analyzing data across different types of research methods (Kongsved et al., 2007; Lazer et al., 2009). The time and ease of online surveys and experiments have been studied in comparison to many traditional face-to-face and telephone research methods (Lallukka et al., 2020; Lindhjem & Navrud, 2011), and self-administered paper-and-pencil questionnaires (Kongsved et al., 2007; Weigold et al., 2013).

Online surveys and experiments may help save time for participants. For more experienced users, taking surveys or participating in a survey using a keyboard, mouse, computer screen, or mobile device may make answering an online questionnaire or navigating an experiment much faster than writing responses by hand or giving their answers to an interviewer who then records them (Nimrod, 2018). However, faster participation times in online surveys and experiments may indicate that respondents are paying less attention to the online stimuli or survey questions and may result in a lower level of data quality (Wenz, 2021). Data from online surveys, experiments, content analyses, and online qualitative approaches can be easily exported into data analysis/analytic programs like SPSS, R, NVivo, and a variety of other tools.

Incorporating Multimedia and Monitoring Participant Behavior

Compared to traditional methods, online survey/experiment platforms allow researchers to include multimedia (e.g., videos, photos, audio recordings, other media) sources, which can serve as stimuli in an online experimental design or to enhance the online survey experience for participants. In terms of online content analysis, mobile applications allow researchers to obtain diverse sources of data unobtrusively (e.g., physical activity tracking using a pedometer or Global Positioning System, non-intrusive biometric data – heart activity or blood pressure, and textual data – linguistic patterns in user writing). In addition, social media platform and mobile application analytics (e.g., number of clicks, likes/dislikes, shares, usage time, etc.) can be conveniently captured and analyzed (Wright et al., 2019).

Overcoming Geographic and Temporal Constraints

Another advantage of online research methods includes access to individuals in distant locations, the ability to reach difficult-to-contact participants, and the convenience of having automated data collection (which reduces researcher time and effort). Survey-based research has been shown to be comparable to mailed surveys in terms of response rates and quality of data (Ibarra et al., 2018; Ramo & Prochaska, 2012; Wright, 2017). Researchers can easily reach participants from all over the United States, including both urban and rural areas using online methods. Similarly, researchers have the ability to use the Web to access participants from all around the world as well as more easily collaborate with colleagues in other countries. However, one caveat is that Internet access and use are not equally distributed worldwide. A substantial digital divide exists between privileged and underprivileged socioeconomic groups and countries (Pullmann et al., 2009). In general, the countries with the greatest Internet access are typically more affluent, better educated, and have a higher gross domestic product (GDP) rate.

Access to Hard-to-Reach Populations

Many people who use the Internet and social media are drawn to specific online groups, communities, and social media platforms based on common interests, and this can be helpful for researchers to access a concentrated number of people who share common interests, beliefs, attitudes, behaviors (Wright, 2005). Scholars have used online methods to study a wide variety of hard-to-reach populations (King et al., 2014; Russomanno et al., 2019). For example, researchers can use online services to help them generate panels for longitudinal surveys or to locate a wider range of individuals who share common interests or characteristics based on their online activity (Beymer et al., 2018; Christenson & Glick, 2013).

In addition, the advent of autocoding certain words, participant profile characteristics, or online messages has allowed researchers to capture large amounts of data for content analysis studies of unique online samples or populations. Big data technologies allow for the analysis of large-scale, rapidly generated (often in real time), and complex sets of data that can be useful in a wide variety of research programs (Gandomi & Haider, 2015). Social media analytics programs can be used to capture day-to-day, micro-level online behaviors, which can offer better ecological validity in terms of tracking behavior compared to traditional laboratory settings. For example, such programs can capture and track users' everyday language, which can be used for natural language analysis to identify certain behaviors (problematic or desirable) that could be used in health interventions and marketing campaigns.

Use of Online Participant Recruitment Services

Online surveys have become increasingly popular in the social and behavioral sciences due to sites such as Amazon's Mechanical Turk or Qualtrics panels. Mechanical Turk is currently a common source for many researchers to conduct both online surveys and experiments (see Christenson & Glick, 2013; Dietrich & Winters, 2015). Mechanical Turk's low cost and recruitment speed make it an excellent method for pretesting, exploratory research, and designs that depend on current events (Christenson & Glick, 2013). Qualtrics panels offer similar advantages as Mechanical Turk, although they are often more expensive. A growing literature has evaluated Mechanical Turk samples in the United States by comparing them to probability samples and/or traditional, in-person convenience samples (Clifford et al., 2015; Levay et al., 2016). Such studies have replicated findings from traditional laboratory experiments using online experimental designs and samples. In short, the data obtained online are as reliable as those obtained via traditional methods.

Self-Administration, Reduced Social Desirability Bias, and Reduced Impact of Researcher Influence on Responses

As with all self-administered surveys, online surveys and experiments may be more convenient for respondents because they can answer the survey at their own pace, whenever and from wherever they choose. The absence of visible interviewers in online surveys creates another important benefit – the reduction of unintentional researcher attribute effect and/or the absence or reduction of researcher non-verbal communication behaviors that might signal socially desirable answers (see Chapter 11 in this volume). Respondents who answer sensitive questions in private are often more open and tend to yield less to socially desirable answers. For example, studies have found that online survey participants tend to be more negative about immigrants (Herwegh & Loosveldt, 2008), more likely to admit to legal offenses (Bronner & Kuijlen, 2007) and unhealthy behaviors (e.g., excessive alcohol consumption; Link & Mokdad, 2005), or illegal drug use (Dietz et al., 2013) compared to traditional survey methods. Similarly, the absence of face-to-face interviewers is beneficial in terms of obtaining responses to sensitive questions (Joinson et al., 2007).

Disadvantages of Online Research Methods

Validity and Participant Privacy Concerns

Online research methods are not always able to reach some elements of the target population. For example, only respondents with Internet access can complete online surveys or participate in online experiments. Many web surveys rely on self-selection of respondents instead of probability sampling, and this can have a negative impact on the quality or generalizability of survey or online experiment results (Kramer et al., 2014; Lefever & Matthiasdottir, 2007). However, weighting adjustment techniques can be used help to reduce selection bias in some cases (Greenacre, 2016). Institutional review boards (IRBs) at most major universities typically have some guidelines regarding the conduct of web-based research, particularly participant confidentiality and privacy issues. However, depending upon the sophistication of the survey design, the IRB may have additional concerns or questions for a researcher to address (see Chapter 2 in this volume).

Limitations of Online Experiments

Despite their great potential, online experiments are often limited to non-interactive directions or decision-making tasks for participants. While platforms like Qualtrics or SurveyMonkey allow researchers to document decision-making behaviors for tasks that participants complete individually, they do not easily permit the use of interactions involving live feedback between participants; in the laboratory, the experimenter can monitor and enforce any restriction of communication between participants to ensure they are completing online surveys or experiments independently. However, this is much more challenging in an online experiment. Moreover, it may be more difficult to reduce interparticipant bias online when drawing participants from the same online community or pool for an online survey or experiment (Edlund et al., 2017).

Best Practices When Using Online Surveys

Over the past two decades, we have seen considerable growth in the area of online survey methodology, particularly in the areas of online survey development and implementation (Dillman, 2000; Greenlaw & Brown-Welty, 2009; Kramer et al., 2014; Lieberman, 2008; Murray et al., 2009; Wright, 2005). The literature regarding online survey methodology has identified and described several concerns, including potential biases and data quality (Eysenbach & Wyatt, 2002; McInroy, 2016; Mullinix et al., 2015) technological issues, and ethical considerations. This body of work has identified a number of best practices that may help researchers to overcome such obstacles when conducting online research.

Non-response from potential online sample members can represent a significant problem for online surveys to the same degree as for traditional paper questionnaires (Coste et al., 2013; Hohwü et al., 2013). However, researchers have identified several factors that appear to increase response rates in online surveys, including personalized email invitations, follow-up reminders, pre-notification of the intent to survey, and simpler/shorter web questionnaire formats (Cook et al., 2000; Galesic & Bosnjak, 2009; see also Chapter 9 in this volume). Other factors that increase response rates include incentives, credible sponsorship of the survey, and multi-modal approaches

(Fan & Yan, 2010). Kaplowitz et al. (2004) found that a web survey application achieved a comparable response rate to a mail hard copy questionnaire when both were preceded by an advance mail notification. In addition, reminder mail notifications had a positive effect on response rate for the online survey compared to a treatment group in which participants only received an email containing a link to the online survey.

Online surveys may present problems due to technical hardware and software issues, while a person is completing the survey or when storing the data, or the refusal of participants to provide information by using non-committal replies (Denscombe, 2009). In online surveys, there is no single response rate. Instead, there are multiple potential methods for calculating a response rate (Mullinix et al., 2015). Another common concern for online surveys is that a single user fills in the same questionnaire multiple times (Mullinix et al., 2015). Multiple methods are available to detect and hopefully prevent or minimize the chance of this occurring (e.g., using cookies or IP [Internet Protocol] analysis; Denscombe, 2009).

Moreover, online surveys can easily take advantage of advancing technology to provide multiple-question formats, direct database connectivity, data-quality checking, customized instrument delivery, and guaranteed confidentiality – all of which can serve to improve the reliability of the data (Mullinix et al., 2015). Online surveys do not appear to compromise the psychometric properties of common quantitative measures (e.g., Likert-type scales, etc.), and participants are typically not less representative of the general population compared to traditional studies (Denissen et al., 2010). Although it may take less time to reach a sufficient sample size using online surveys, many responses from online participants may be left blank (unless the researcher requires participants to complete every question). As a result, what may look to be an initial sample of 300 on Qualtrics may have large numbers of unusable responses from participants. I recommend to over-sample by 20–30% responses over the initial target response rate goal to account for this.

Another problem that can occur with longitudinal online surveys is participant attribution. However, studies suggest that attrition in online longitudinal surveys does not differ from traditional surveys (Fan & Yan, 2010). Moreover, automated email reminders are a cheap and convenient way to reduce attrition (Fan & Yan, 2010). While most online surveys tend to be cross-sectional, many researchers have conducted such longitudinal surveys successfully (Leach et al., 2016; Valkenburg & Peter, 2007). It is important in longitudinal studies for researchers to record and analyze participant attrition.

Many individuals do not access online surveys via a desktop or laptop computer (Antoun et al., 2017). Instead, they use a smartphone (or other mobile device) for all their electronic communication needs. For example, studies have found that over half of US adults own a smart phone, and many (especially younger individuals) use this device instead of a computer (Antoun et al., 2017). Accessing an online survey via a smartphone requires utilizing formats that require very little space (Callegaro, 2010). Some traditional question formats (e.g., matrices) may be more difficult to use when a significant portion of one's sample relies only on such devices.

Online survey researchers have recommended posting an open invitation link within an online community or sending out invitations to the entire target population (Murray et al., 2009). This has been found to increase response rates from online community members (Murray et al., 2009). As smartphones have increasingly become the norm in terms of everyday communication, respondents may experience increased willingness to complete online surveys that are tailored for smartphone use (Nayak & Narayan, 2019).

Best Practices Using Online Experimental Designs

To conduct online experiments, researchers typically need some type of browser-based experimental platform, a server to host the experiment, and a participant recruitment tool (Grootswagers, 2020). An experiment needs to run in a web browser, so it must be programmed in a browser-compatible programming language (e.g., JavaScript; De Leeuw, 2015). For those researchers who are less technically inclined, popular survey platforms, such as Qualtrics, allow scholars to create and host relatively complex experiments within the platform. Moreover, Qualtrics, and other online survey tools, offer participant recruitment options for researchers (for a fee). Such services may help researchers bypass learning a programming language, find a server to host the experiment, and have a means for obtaining a representative sample (Mutz, 2011). However, several online survey platforms and services are available to researchers depending on the type of online experimental studies they wish to conduct. These platforms and services vary in terms of the features that they specifically offer to researchers; the cost of using them may increase depending upon the needs/requirements of the study. For example, some platforms offer a complete experiment-hosting infrastructure, such as Testable, Inquisit, and Gorilla (Anwyl-Irvine et al., 2020). However, there are also several free and open-source experiment builders that can export experiments as browser-compatible JavaScript code (e.g., Psychopy; Peirce et al., 2019).

A growing body of studies has shown that online experiments can yield results comparable to those obtained in conventional laboratory settings (Casler et al., 2013; Dandurand et al., 2008; Hilbig, 2016). Online experiments may provide a critical baseline of comparison for researchers when running multiple studies. For example, recruiting multiple samples for online experiments may help researchers procure many homogenous samples that would be more time consuming or expensive face-to-face. Online experiments also accurately replicate the findings from behavioral experiments that rely on reaction time measurement and learning tasks with complex instructions (Barnhoorn et al., 2015).

Larger and diverse samples also provide the ability to test different populations as moderating variables, which may expand researchers' ability to assess the influence of cultural factors and location when testing theoretical models. In addition, recruiting samples online for online experiments allows for some degree of increased diversity/representativeness compared to the heavy reliance on undergraduate student samples in traditional laboratory research in university settings (Parigi et al., 2017).

To conduct studies with larger and more diverse samples, researchers have developed and evaluated alternative ways to recruit participants, such as through Mechanical Turk or similar platforms (Radford et al., 2016). Finding and maintaining an active pool of potential participants is the main advantage of such services, although they can be expensive. Compared to traditional laboratory experiments, online studies often offer faster and more effortless participant recruitment (Parigi et al., 2017). The convenience and lower cost of conducting experiments online and the use of online recruiting services has resulted in numerous large-scale studies comparing multiple demographic groups, ages, languages, and countries (Parigi et al., 2017; Woo et al., 2015; Zhang et al., 2018).

Another prominent advantage of running experimental studies online lies in its efficiency. It is possible to collect responses from hundreds of participants within hours due to the potential of worldwide sampling. Platforms and services allow for a large number of participants to be tested simultaneously; this would not be possible in a face-to-face laboratory-based setting (Zhang et al., 2018. Online experiments are not restricted to office hours or teaching schedules, do not require hard resources, and do not require an in-person presence for participants or researchers (Grootswagers, 2020).

However, a major concern when conducting experiments online is data quality (Grootswagers, 2020). Some concerns (e.g., motivation, distractions, stimulus timing) can be alleviated with an appropriate design and incentive strategy. Online experiments only work for some stimulus modalities. While the online approach is well suited for experiments consisting of visual stimuli and keyboard or mouse responses, other paradigms are harder or impossible to move online. Another limitation is the lack of experimental control. For example, there is no way to know the participant's distance from the screen (Grootswagers, 2020). This makes it impossible to control the visual angle of stimuli – a limiting factor for some experiments. It is also hard to test whether participants are paying attention to the experiment. Another problem that may affect data quality is participant attrition in online experiments. Unlike laboratory studies, participants may drop out at rates of up to 69%. In a dropout analysis of 88 local studies, Zhou and Fishbach (2016) found that 20% had a dropout rate of over 30%.

To facilitate participation, online experimenters need to be very thorough when creating experimental instructions – so that they can appear as "stand alone" directions – since it is unlikely that a specific participant will be able to interact with the researcher as he or she completes the online experimental tasks (due to different time zones, etc.). It is also important that the instructions are comprehensible by people of a wider age range, cultures, and socio-economic backgrounds (Crump et al., 2013; Reimers & Stewart, 2015). Using a pictorial step-by-step set of instructions may lead to fewer misunderstandings compared to a single page of text. Researchers may want to ensure that the instructions for the experiment stay on the screen for some time before continuation is allowed or an instruction check is added. However, there is no guarantee that this extended time will lead to reading and understanding the instructions. Researchers do have the ability to monitor some functions to check that participants stayed on track during the experiment. For example, it is possible to monitor how often the browser tab running the experiment was minimized during the experiment (Gureckis et al., 2016). Finally, online experimental studies should be short. Fatigue may occur with longer experiments, increasing the possibility of distractions within a participant's personal environment (e.g., children, etc.) (Hamby & Taylor, 2016).

Best Practices Conducting Online Content Analyses

Social media and other digital content are widely accessible, constantly added to, and available in an easy-to-access electronic format (compared to more traditional texts). Using the information obtained from content analyses, researchers have gained valuable insights into the beliefs, attitudes, and perceptions of people who use online communities, mobile applications, microblogs (i.e., Twitter), and a wide variety of other digital resources that can be accessed via the Web (Chew & Eysenbach, 2010; De Wever et al., 2006; Lee et al., 2014).

Content analysis is a systematic technique for unobtrusively coding symbolic content (text, images, etc.) found online, especially structural features (e.g., message length, distribution of certain text or image components) and semantic themes (Krippendorf, 2018). Although the primary use of content analysis is to identify and describe patterns in manifest content, the technique can also be used for making inferences about intentions and effects (Holsti, 1969; Krippendorf, 2018). Establishing a careful set of coding criteria and use of a codebook for training coders is a hallmark of content analysis studies, although researchers should also be open to emergent phenomena that may surface in online settings as well. Moreover, the dynamic nature and sheer number of units of Internet analysis can make random sampling infeasible in many cases (Riffe et al., 2019; Schneider & Foot, 2004).

The abundance of web pages and their diversity of form and function (as well as the unprecedented ease with which content can be collected and analyzed using automated tools) provide seemingly endless opportunities for research. The term "big data" has emerged in recent years to describe the volume of information produced by online users, made possible by the growing ubiquity of mobile devices, tracking tools, always-on sensors, and cheap computing storage (Manyika et al., 2011). Technological advances have made it easier than ever to harness, organize, and scrutinize large repositories of digital information. Advances in computational techniques for large-scale data analysis, that once required supercomputers, now can be conducted on a desktop computer (Manovich, 2012). This development has created exciting opportunities for computational approaches to research (Lazer et al., 2009). For example, the dramatic growth of social network sites has provided a massive amount of data that reflect new media activities (e.g., tweets, status updates, shares). This allows researchers to explore novel means of analyzing media content, as they use computational methods to assemble, filter, and interpret content that is created via Web 2.0 around a particular topic or event (Riffe et al., 2019).

McMillan (2000) identified a number of challenges to applying content analysis to the Web, including difficulties obtaining a representative sample, defining the unit of analysis, and ensuring that coders are presented with the same content for purposes of reliability. However, in many cases, online content analysis may only require minor adaptations to traditional approaches of content analysis, such as using lists to help generate sampling frames and using software to capture website content (e.g., Radian6). Newer computational methods offer the potential for overcoming some of the sampling and coding limitations of traditional content analysis. Algorithmic techniques can be used to reduce a vast body of data into smaller pools of data for specialized analyses; web analytics programs can help researchers access data from other online services (e.g., digital companies such as TripAdvisor, Yelp, etc.). These features can generate giant volumes of content that can be analyzed by researchers, and give insights into consumer perceptions and behaviors (Gandomi & Haider, 2015).

Finally, social network analysis is a form of content analysis (Krippendorf, 2018; Williams & Shepherd, 2017). Social network analysis can be used to analyze networks of ties (e.g., as constituted by communication or transaction) between nodes (e.g., people, institutions, etc.). Social network analysis is also well suited for analyzing patterns of relationships on social media and other digital platforms (Pfeil & Zaphiris, 2009; Takahashi et al., 2009). Such information may be important to researchers interested in concepts like social influence or social support. Understanding online social networks can provide researchers with important insights into how messages (legitimate information and misinformation) are disseminated via online social networks.

Best Practices with Online Qualitative Research Approaches

The Internet potentially provides qualitative researchers with a variety of new approaches for conducting research, new venues for social research, and new means for understanding the way social realities are constructed and reproduced through human interaction (Hallett & Barber, 2014). Internet qualitative research methods can be broadly defined as methods that are used to collect qualitative data for interviews, observation, and/or document analyses (Markham, 2005). However, given the wide range of qualitative approaches to studying online spaces, we will focus only on more general considerations when conducing online ethnography approaches and online interviews/focus groups.

Online Ethnographic Approaches

Ethnographic approaches to online research are quite popular and often go by the names of “virtual ethnography” (Hine, 2000) and “netnography” (Kozinets, 2002). Online qualitative researchers have argued that observing online social phenomena from a qualitative standpoint is important in terms of exploring how social realities are constructed through online interactions and social processes (Larsen, 2008; Murthy, 2011). Online ethnographers often use approaches that are similar to traditional face-to-face ethnographic data collection, including observation, interviews with key informants (i.e., members of the online culture or community being observed), and qualitative textual analysis (see Bortree, 2005; Johnson & Humphry, 2012; Manninen, 2017; Wang & Sandner, 2019). Online ethnography may allow for more detached research observation that may reduce the researcher's input and bias compared to face-to-face settings, and researchers can collect data unobtrusively by simply observing and recording behaviors within a variety of online settings. However, researchers who observe online groups or communities without participating have sometimes been referred to as "lurkers," a practice that has been condemned by some scholars (Bell, 2006).

Online Ethnographic Interviews

An alternative approach for researchers conducting online research is to actively talk to participants, as opposed to observing online content or already occurring conversations in online spaces (Kozinets, 2010). This often takes the form of qualitative interviews with key informants who can provide researchers with important emic, or insider, perspectives of a wide range of online community beliefs and behaviors (Hoare et al., 2013; Salmons, 2014). Interviews with key informants who have extensive cultural knowledge of these online communities may yield important insights for researchers. In some cases, these insights may inform intervention strategies or provide explanations for certain interactions and behaviors on the Web (e.g., why group members have a certain belief or how they influence one another).

However, online platforms allow individuals to create their own self-presentations. Research on computer-mediated communication has consistently documented people's ability to strategically alter information or selectively reveal only certain aspects about their identity online (see Walther, 2007; Walther & Burgoon, 1992). As a result, researchers who use qualitative methods need to be cognizant of how features of the computer-mediated environment influence self-presentation. For example, both asynchronous and synchronous computer-mediated communication formats allow people to engage in selective self-presentation (e.g., presenting certain aspects of oneself while hiding others) in ways that would be difficult or impossible in the face-to-face world due to the reduced non-verbal cues in online communication. The success of qualitative interviewing and ethnographic access often depend on the relationships of trust a researcher can build, their access to participants' social worlds, and how much they "get" from participants.

Online Focus Groups

Focus groups differ from interviews with individuals by bringing together people with mutual characteristics or interests to offer individual and collective insights into particular topics (Kenny, 2005). Online focus groups can be conducted in real time on a variety of platforms (e.g., Skype, Zoom, etc.) and are comparable to conversational interactions seen in face-to-face focus groups (Fox et al., 2007); they can also be asynchronous – using "static" text-based communication (e.g., emailing questions to individual focus group members; Kenny, 2005). Online focus groups can bring together geographically distant individuals and groups in web-based settings, offer practical advantages (e.g., avoiding costly and difficult transcription of focus group conversations), and facilitate greater participation and disclosure for users who may be more comfortable interacting with a researcher online. However, similar to problems with individual interviews online, online focus groups may be affected by features of the computer-mediated environment (e.g., reduced or absent nonverbal cues, etc.). Such factors may influence group dynamics, group members' willingness to communicate about certain topics, and how much moderators can facilitate and control online focus group discussions. Due to the egalitarian nature of online communication, another concern is that characteristics of computer-mediated communication may undermine the position of the researcher as a professional authority.

Qualitative Textual Analysis

Finally, online researchers often make use of online artifacts related to a particular online community of interest (e.g., social media conversations, website content, etc.). These can be analyzed qualitatively to access and identify themes regarding online phenomena that might not be possible to study through observation, interviews, or focus groups. Such approaches in online ethnography are often used in conjunction with observations and interviews. Qualitative data analysis software (e.g., NVivo) allow qualitative researchers to conveniently import online cultural artifacts (e.g., web content, online conversations) as well as field notes and interview transcripts from Word documents into the program. NVivo facilitates the coding of qualitative data from these varied sources, and it allows qualitative researchers to identify common themes and exemplars from the qualitative data more easily and cheaply than traditional ethnographic data analysis methods.

Other Considerations

The Internet provides qualitative researchers with access to otherwise hard-to-reach populations (e.g., members of online support groups). Other research methods may not be appropriate for studying these types of populations as members may be hesitant to complete an online survey or participate in an online experiment if they are cognizant that they hold minority or unpopular views regarding certain issues (Kraut et al., 2004). Interviews with key informants who have extensive cultural knowledge of these online communities may yield important insights for researchers.

There are several ethical issues that online qualitative researchers may face when studying online phenomena (see Chapter 2 in this volume). Given the variability and changing nature of online spaces, it is important for researchers to consider how they are conducting research on the Internet and whether they need to revisit or pay extra attention to certain aspects of the research process (e.g., obtaining consent multiple times, as the membership composition of online communities may change over time). In addition, social media groups and websites can vary in terms of privacy and third-party use, including whether they allow researchers to observe posts or conversations by participants. Researchers need to be aware that users of such sites have varied levels of knowledge regarding how their uploaded content is used or accessed. In some cases, the very act of being known to others in an online community as a study participant might carry risks, and researchers are expected to protect participant confidentiality.

Part of the expectation of informed consent for participation in research is that participants can choose what information to disclose, how to present themselves, and what level of access to allow the researcher. We expect participants will regulate what information they will disclose to a researcher in online settings. However, researchers may be tempted to conduct searches of social media posts and other online sources to learn more about a participant. While such searches may shed light on important characteristics of participants, researchers need to be aware of the limits of informed consent as well as protecting participant privacy (especially in research articles and other publication outlets). When considering such ethical issues, it is important for researchers to provide transparent accounts on how they accessed data online and what ethical protocols they followed.

Conclusion

The purpose of this chapter was to examine advantages, disadvantages, and best practices identified by online researchers when using four common online research methods: online surveys, online experiments, online content analysis, and qualitative approaches to studying online populations. In many cases, online research methods raise many of the same concerns for researchers as traditional face-to-face research methods. In some cases, online research methods present new challenges and opportunities for researchers, such as the ability of participants to more easily engage in selective self-presentation or shift their identities from one online platform to another.

Online research methods are constantly changing as new data capture and analysis techniques and software allow researchers new ways to more efficiently conduct studies. Experimental design platforms like Gorilla have provided opportunities for researchers to control more variables in online experiments; programs like Radian6 and Python allow researchers to more easily capture and organize online content; and autocoding of content in online content analyses and software allow qualitative researchers to more easily integrate field notes, interview transcripts, and qualitative texts for data analysis. Mixed online research methods allow opportunities for researchers to compliment the strengths of various online research methods as well as help offset some of the limitations of individual methods. Researchers should continue to use online research methods and document their various strengths and limitations. While online research methods may not be compatible with all types of research agendas, they can be useful in terms of pilot-testing data or refining studies prior to moving them into a more traditional research setting. In the future, scholars should continue to find innovative ways to increase the sophistication and refinement of current online research methods.

References

Antoun, C., Couper, M. P., & Conrad, F. G. (2017). Effects of mobile versus PC web on survey response quality: A crossover experiment in a probability web panel. *Public Opinion Quarterly*, *81*(S1), 280–306. https://doi.org/10.1093/poq/nfw088

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Aristeidou, M., Scanlon, E., & Sharples, M. (2017). Profiles of engagement in online communities of citizen science participation. *Computers in Human Behavior*, *74*, 246–256. https://doi.org/10.1016/j.chb.2017.04.044

Armstrong, B., Reynolds, C., Bridge, G., et al. (2020). How does citizen science compare to online survey panels? A comparison of food knowledge and perceptions between the Zooniverse, Prolific and Qualtrics UK panels. *Frontiers in Sustainable Food Systems*, *4*, 306. https://doi.org/10.3389/fsufs.2020.575021

Babbie, E. R. (2020). *The Practice of Social Research*. Cengage Learning.

Barnhoorn, J. S., Haasnoot, E., Bocanegra, B. R., & van Steenbergen, H. (2015). QRTEngine: An easy solution for running online reaction time experiments using Qualtrics. *Behavior Research Methods*, *47*(4), 918–929. https://doi.org/10.3758/s13428-014-0530-7

Bell, D. (2006). *An Introduction to Cybercultures*. Routledge.

Beymer, M. R., Holloway, I. W., & Grov, C. (2018). Comparing self-reported demographic and sexual behavioral factors among men who have sex with men recruited through Mechanical Turk, Qualtrics, and a HIV/STI clinic-based sample: Implications for researchers and providers. *Archives of sexual behavior*, *47*(1), 133–142. https://doi.org/10.1007/s10508-016-0932-y

Bortree, D. S. (2005). Presentation of self on the Web: An ethnographic study of teenage girls' weblogs. *Education, Communication & Information*, *5*(1), 25–39. https://doi.org/10.1080/14636310500061102

Bronner, F. & Kuijlen, T. (2007). The live or digital interviewer: A comparison between CASI, CAPI and CATI with respect to differences in response behaviour. *International Journal of Market Research*, *49*(2), 167–190. https://doi.org/10.1177/147078530704900204

Callegaro, M. (2010). Do you know which device your respondent has used to take your online survey. *Survey Practice*, *3*(6), 1–12.

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, *29*(6), 2156–2160. https://doi.org/10.1016/j.chb.2013.05.009

Chew, C. & Eysenbach, G. (2010). Pandemics in the age of Twitter: Content analysis of Tweets during the 2009 H1N1 outbreak. *PloS One*, *5*(11), e14118. https://doi.org/10.1371/journal.pone.0014118

Christenson, D. P. & Glick, D. M. (2013). Crowdsourcing panel studies and real-time experiments in MTurk. *The Political Methodologist*, *20*(2), 27–32.

Clifford, S., Jewell, R. M., & Waggoner, P. D. (2015). Are samples drawn from Mechanical Turk valid for research on political ideology? *Research & Politics*, *2*(4). https://doi.org/10.1177/2053168015622072

Cook, C., Heath, F., & Thompson, R. L. (2000). A meta-analysis of response rates in web-or Internet-based surveys. *Educational and Psychological Measurement*, *60*(6), 821–836. https://doi.org/10.1177/00131640021970934

Coste, J., Quinquis, L., Audureau, E., & Pouchot, J. (2013). Non response, incomplete and inconsistent responses to self-administered health-related quality of life measures in the general population: Patterns, determinants and impact on the validity of estimates – a population-based study in France using the MOS SF-36. *Health and Quality of Life Outcomes*, *11*(1), 1–15. https://doi.org/10.1186/1477-7525-11-44

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS One*, *8*(3), e57410. https://doi.org/10.1371/journal.pone.0057410

Dandurand, F., Shultz, T. R., & Onishi, K. H. (2008). Comparing online and lab methods in a problem-solving experiment. *Behavior Research Methods*, *40*(2), 428–434. https://doi.org/10.3758/BRM.40.2.428

De Wever, B., Schellens, T., Valcke, M., & Van Keer, H. (2006). Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review. *Computers & Education*, *46*(1), 6–28. https://doi.org/10.1016/j.compedu.2005.04.005

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Denissen, J. J., Neumann, L., & Van Zalk, M. (2010). How the Internet is changing the implementation of traditional research methods, people's daily lives, and the way in which developmental scientists conduct research. *International Journal of Behavioral Development*, *34*(6), 564–575. https://doi.org/10.1177/0165025410383746

Denscombe, M. (2009). Item non-response rates: A comparison of online and paper questionnaires. *International Journal of Social Research Methodology*, *12*(4), 281–291. https://doi.org/10.1080/13645570802054706

Dietrich, S. & Winters, M. S. (2015). Foreign aid and government legitimacy. *Journal of Experimental Political Science*, 2(2), 164–171. https://doi.org/10.1017/XPS.2014.31

Dietz, P., Striegel, H., Franke, A. G., et al. (2013). Randomized response estimates for the 12-month prevalence of cognitive-enhancing drug use in university students. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, *33*(1), 44–50. https://doi.org/10.1002/phar.1166

Dillman, D. A. (2000). Procedures for conducting government-sponsored establishment surveys: Comparisons of the total design method (TDM), a traditional cost-compensation model, and tailored design. In Proceedings of American Statistical Association, Second International Conference on Establishment Surveys, Buffalo, New York, June 17–21 (pp. 343–352).

Edlund, J. E., Lange, K. M., Sevene, A. M., et al. (2017). Participant crosstalk: Issues when using the Mechanical Turk. *Tutorials in Quantitative Methods for Psychology*, *13*(3), 174–182. https://doi.org/10.20982/tqmp.13.3.p174

Eysenbach, G. & Till, J. E. (2001). Ethical issues in qualitative research on internet communities. *BMJ*, *323*(7321), 1103–1105. https://doi.org/10.1136/bmj.323.7321.1103

Eysenbach, G. & Wyatt, J. (2002). Using the Internet for surveys and health research. *Journal of Medical Internet Research*, *4*(2), e13. https://doi.org/10.2196/jmir.4.2.e13

Fan, W. & Yan, Z. (2010). Factors affecting response rates of the web survey: A systematic review. *Computers in Human Behavior*, *26*(2), 132–139. https://doi.org/10.1016/j.chb.2009.10.015

Fox, F. E., Morris, M., & Rumsey, N. (2007). Doing synchronous online focus groups with young people: Methodological reflections. *Qualitative Health Research*, *17*(4), 539–547. https://doi.org/10.1177/1049732306298754

Galesic, M. & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, *73*(2), 349–360. https://doi.org/10.1093/poq/nfp031

Gandomi, A. & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Greenacre, Z. A. (2016). The importance of selection bias in Internet surveys. *Open Journal of Statistics*, *6*(03), 397. https://doi.org/10.4236/ojs.2016.63035

Greenlaw, C. & Brown-Welty, S. (2009). A comparison of web-based and paper-based survey methods: Testing assumptions of survey mode and response cost. *Evaluation Review*, *33*(5), 464–480. https://doi.org/10.1177/0193841X09340214

Grootswagers, T. (2020). A primer on running human behavioural experiments online. *Behavior Research Methods*, *1*(4), 2283–2286. https://doi.org/10.3758/s13428-020-01395-3

Gureckis, T. M., Martin, J., McDonnell, J., (2016). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, *48*(3), 829–842. https://doi.org/10.3758/s13428-015-0642-8

Hall, M. G., Grummon, A. H., Lazard, A. J., Maynard, O. M., & Taillie, L. S. (2020). Reactions to graphic and text health warnings for cigarettes, sugar-sweetened beverages, and alcohol: An online randomized experiment of US adults. *Preventive Medicine*, *137*, 106120. https://doi.org/10.1016/j.ypmed.2020.106120

Hallett, R. E. & Barber, K. (2014). Ethnographic research in a cyber era. *Journal of Contemporary Ethnography*, *43*(3), 306–330. https://doi.org/10.1177/0891241613497749

Hamby, T. & Taylor, W. (2016). Survey satisficing inflates reliability and validity measures: An experimental comparison of college and Amazon Mechanical Turk samples. *Educational and Psychological Measurement*, *76*(6), 912–932. https://doi.org/10.1177/0013164415627349

Heerwegh, D. & Loosveldt, G. (2008). Face-to-face versus web surveying in a high-internet-coverage population: Differences in response quality. *Public Opinion Quarterly*, *72*(5), 836–846. https://doi.org/10.1093/poq/nfn045

Hilbig, B. E. (2016). Reaction time effects in lab-versus web-based research: Experimental evidence. *Behavior Research Methods*, *48*(4), 1718–1724. https://doi.org/10.3758/s13428-015-0678-9

Hine, C. (2000). *Virtual Ethnography*. SAGE Publications.

Hoare, K. J., Buetow, S., Mills, J., & Francis, K. (2013). Using an emic and etic ethnographic technique in a grounded theory study of information use by practice nurses in New Zealand. *Journal of Research in Nursing*, *18*(8), 720–731. https://doi.org/10.1177/1744987111434190

Hohwü, L., Lyshol, H., Gissler, M., et al. (2013). Web-based versus traditional paper questionnaires: A mixed-mode survey with a Nordic perspective. *Journal of Medical Internet Research*, *15*(8), e173. https://doi.org/10.2196/jmir.2595

Holsti, O. R. (1969). *Content Analysis for the Social Sciences and Humanities*. Addison-Wesley.

Ibarra, J. L., Agas, J. M., Lee, M., Pan, J. L., & Buttenheim, A. M. (2018). Comparison of online survey recruitment platforms for hard-to-reach pregnant smoking populations: Feasibility study. *JMIR Research Protocols*, *7*(4), e8071. https://doi.org/10.2196/resprot.8071

Johnson, N. F. & Humphry, N. (2012). The Teenage Expertise Network (TEN): An online ethnographic approach. *International Journal of Qualitative Studies in Education*, *25*(6), 723–739. https://doi.org/10.1080/09518398.2011.590160

Joinson, A. N., Woodley, A., & Reips, U. D. (2007). Personalization, authentication and self-disclosure in self-administered Internet surveys. *Computers in Human Behavior*, *23*(1), 275–285. https://doi.org/10.1016/j.chb.2004.10.012

Kaplowitz, M. D., Hadlock, T. D., & Levine, R. (2004). A comparison of web and mail survey response rates. *Public Opinion Quarterly*, *68*(1), 94–101. https://doi.org/10.1093/poq/nfh006

Kenny, A. J. (2005). Interaction in cyberspace: An online focus group. *Journal of Advanced Nursing*, *49*(4), 414–422. https://doi.org/10.1111/j.1365-2648.2004.03305.x

King, D. B., O'Rourke, N., & DeLongis, A. (2014). Social media recruitment and online data collection: A beginner's guide and best practices for accessing low-prevalence and hard-to-reach populations. *Canadian Psychology/Psychologie Canadienne*, *55*(4), 240. https://doi.org/10.1037/a0038087

Kongsved, S. M., Basnov, M., Holm-Christensen, K., & Hjollund, N. H. (2007). Response rate and completeness of questionnaires: A randomized study of Internet versus paper-and-pencil versions. *Journal of Medical Internet Research*, *9*(3), e25. https://doi.org/10.2196/jmir.9.3.e25

Kozinets, R. V. (2002). The field behind the screen: Using netnography for marketing research in online communities. *Journal of Marketing Research*, *39*, 61–72. https://doi.org/10.1509/jmkr.39.1.61.18935

Kozinets, R. V. (2010). *Netnography: Doing Ethnographic Research Online*. SAGE Publications.

Kramer, J., Rubin, A., Coster, W., et al. (2014). Strategies to address participant misrepresentation for eligibility in Web-based research. *International Journal of Methods in Psychiatric Research*, 23(1), 120–129. https://doi.org/10.1002/mpr.1415

Kraut, R., Olson, J., Banaji, M., et al. (2004). Psychological research online: Report of Board of Scientific Affairs' Advisory Group on the Conduct of Research on the Internet. *American Psychologist*, *59*(2), 105. https://doi.org/10.1037/0003-066X.59.2.105

Krippendorff, K. (2018). *Content Analysis: An Introduction to Its Methodology*. SAGE Publications.

Lallukka, T., Pietiläinen, O., Jäppinen, S., et al. (2020). Factors associated with health survey response among young employees: A register-based study using online, mailed and telephone interview data collection methods. *BMC Public Health*, *20*(1), 184. https://doi.org/10.1186/s12889-020-8241-8

Larsen, M. C. (2008). Understanding social networking: On young people's construction and co-construction of identity online. *Online Networking: Connecting People*. Icfai University Press.

Lazer, D., Pentland, A., Adamic, L., et al. (2009). Social science. Computational social science. *Science*, *323*(5915), 721–723. https://doi.org/10.1126/science.1167742

Leach, M. J., Hofmeyer, A., & Bobridge, A. (2016). The impact of research education on student nurse attitude, skill and uptake of evidence-based practice: A descriptive longitudinal survey. *Journal of Clinical Nursing*, *25*(1–2), 194–203. https://doi.org/10.1111/jocn.13103

Lee, H., Wright, K. B., O'Connor, M., & Wombacher, K. (2014). Framing medical tourism: An analysis of persuasive appeals, risks and benefits, and new media features of medical tourism broker websites. *Health Communication*, *29*(7), 637–645. https://doi.org/10.1080/10410236.2013.794412

Lefever, S., Dal, M. & Matthiasdottir, A. (2007). Online data collection in academic research: Advantages and limitations. *British Journal of Educational Technology*, *38*(4), 574–582. https://doi.org/10.1111/j.1467-8535.2006.00638.x

Levay, K. E., Freese, J., & Druckman, J. N. (2016). The demographic and political composition of Mechanical Turk samples. *Sage Open*, *6*(1), 2158244016636433. doi: https://doi.org/10.1177/2158244016636433

Lieberman, D. Z. (2008). Evaluation of the stability and validity of participant samples recruited over the Internet. *Cyberpsychology & Behavior*, *11*(6), 743–745. https://doi.org/10.1089/cpb.2007.0254

Lindhjem, H. & Navrud, S. (2011). Are Internet surveys an alternative to face-to-face interviews in contingent valuation? *Ecological Economics*, *70*(9), 1628–1637. https://doi.org/10.1016/j.ecolecon.2011.04.002

Link, M. W. & Mokdad, A. H. (2005). Effects of survey mode on self-reports of adult alcohol consumption: A comparison of mail, web and telephone approaches. *Journal of Studies on Alcohol*, *66*(2), 239–245. https://doi.org/10.15288/jsa.2005.66.239

Manninen, V. J. (2017). Sourcing practices in online journalism: an ethnographic study of the formation of trust in and the use of journalistic sources. *Journal of Media Practice*, *18*(2–3), 212–228. https://doi.org/10.1080/14682753.2017.1375252

Manovich, L. (2012). How to compare one million images? In D. M. Berry (ed.), *Understanding Digital Humanities* (pp. 249–278). Palgrave Macmillan.

Manyika, J., Chui, M., Brown, B., et al. (2011). *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. McKinsey Global Institute.

Markham, A. N. (2005). The methods, politics, and ethics of representation in online ethnography. In *The SAGE Handbook of Qualitative Research*. SAGE Publications.

Mason, W. & Watts, D. J. (2012). Collaborative learning in networks. *Proceedings of the National Academy of Sciences*, *109*(3), 764–769. https://doi.org/10.1073/pnas.1110069108

McInroy, L. B. (2016). Pitfalls, potentials, and ethics of online survey research: LGBTQ and other marginalized and hard-to-access youths. *Social Work Research*, *40*(2), 83–94. https://doi.org/10.1093/swr/svw005

McMillan, S. J. (2000). The microscope and the moving target: The challenge of applying content analysis to the World Wide Web. *Journalism & Mass Communication Quarterly*, *77*(1), 80–98. https://doi.org/10.1177/107769900007700107

Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science*, *2*(2), 109–138. https://doi.org/10.1017/XPS.2015.19

Murray, E., Khadjesari, Z., White, I., et al. (2009). Methodological challenges in online trials. *Journal of Medical Internet Research*, *11*(2), e9. https://doi.org/10.2196/jmir.1052

Murthy, D. (2011). Emergent digital ethnographic methods for social research. In S. N. Hesse-Biber (ed.), *Handbook of Emergent Technologies in Social Research* (pp. 158–179). Oxford University Press.

Mutz, D. C. (2011). *Population-Based Survey Experiments*. Princeton University Press.

Nayak, M. S. D. P. & Narayan, K. A. (2019). Strengths and weakness of online surveys. *IOSR Journal of Humanities and Social Science*, *24*(5), 31–38. doi: 10.9790/0837-2405053138

Nimrod, G. (2018). Technophobia among older Internet users. *Educational Gerontology*, *44*(2–3), 148–162. https://doi.org/10.1080/03601277.2018.1428145

Parigi, P., Santana, J. J., & Cook, K. S. (2017). Online field experiments: Studying social interactions in context. *Social Psychology Quarterly*, *80*(1), 1–19. https://doi.org/10.1177/0190272516680842

Pechey, R. & Marteau, T. M. (2018). Availability of healthier vs. less healthy food and food choice: An online experiment. *BMC Public Health*, *18*(1), 1–11. https://doi.org/10.1186/s12889-018-6112-3

Peirce, J., Gray, J. R., Simpson, S., et al. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Pfeil, U. & Zaphiris, P. (2009). Investigating social network patterns within an empathic online community for older people. *Computers in Human Behavior*, *25*(5), 1139–1155. https://doi.org/10.1016/j.chb.2009.05.001

Pullmann, H., Allik, J., & Realo, A. (2009). Global self-esteem across the life span: A cross-sectional comparison between representative and self-selected Internet samples. *Experimental Aging Research*, *35*(1), 20–44. https://doi.org/10.1080/03610730802544708

Radford, J., Pilny, A., Reichelmann, A., et al. (2016). Volunteer science: An online laboratory for experiments in social psychology. *Social Psychology Quarterly*, *79*(4), 376–396. https://doi.org/10.1177/0190272516675866

Ramo, D. E. & Prochaska, J. J. (2012). Broad reach and targeted recruitment using Facebook for an online survey of young adult substance use. *Journal of Medical Internet Research*, *14*(1), e28. https://doi.org/10.2196/jmir.1878

Rains, S. A., Peterson, E. B., & Wright, K. B. (2015). Communicating social support in computer-mediated contexts: A meta-analytic review of content analyses examining support messages shared online among individuals coping with illness. *Communication Monographs*, *82*(4), 403–430. https://doi.org/10.1080/03637751.2015.1019530

Reimers, S. & Stewart, N. (2015). Presentation and response timing accuracy in Adobe Flash and HTML5/JavaScript Web experiments. *Behavior Research Methods*, *47*(2), 309–327. https://doi.org/10.3758/s13428-014-0471-1

Riffe, D., Lacy, S., Fico, F., & Watson, B. (2019). *Analyzing Media Messages: Using Quantitative Content Analysis in Research*. Routledge.

Russomanno, J., Patterson, J. G., & Tree, J. M. J. (2019). Social media recruitment of marginalized, hard-to-reach populations: Development of recruitment and monitoring guidelines. *JMIR Public Health and Surveillance*, *5*(4), e14886. https://doi.org/10.2196/14886

Salmons, J. (2014). *Qualitative Online Interviews: Strategies, Design, and Skills*. SAGE Publications.

Schneider, S. M. & Foot, K. A. (2004). The Web as an object of study. *New Media & Society*, *6*(1), 114–122. https://doi.org/10.1177/1461444804039912

Shen, G. C. C., Chiou, J. S., Hsiao, C. H., Wang, C. H., & Li, H. N. (2016). Effective marketing communication via social networking site: The moderating role of the social tie. *Journal of Business Research*, *69*(6), 2265–2270. https://doi.org/10.1016/j.jbusres.2015.12.040

Simmons, A. D. & Bobo, L. D. (2015). Can non-full-probability internet surveys yield useful data? A comparison with full-probability face-to-face surveys in the domain of race and social inequality attitudes. *Sociological Methodology*, *45*(1), 357–387. https://doi.org/10.1177/0081175015570096

Skitka, L. J. & Sargis, E. G. (2006). The Internet as psychological laboratory. *Annual Review of Psychology*, 57, 529–555. https://doi.org/10.1146/annurev.psych.57.102904.190048

Stern, M. J., Bilgen, I., & Dillman, D. A. (2014). The state of survey methodology: Challenges, dilemmas, and new frontiers in the era of the tailored design. *Field Methods*, *26*(3), 284–301. doi: https://doi.org/10.1177/1525822X13519561

Takahashi, Y., Uchida, C., Miyaki, K., et al. (2009). Potential benefits and harms of a peer support social network service on the Internet for people with depressive tendencies: Qualitative content analysis and social network analysis. *Journal of Medical Internet Research*, *11*(3), e29. https://doi.org/10.2196/jmir.1142

Valkenburg, P. M. & Peter, J. (2007). Online communication and adolescent well-being: Testing the stimulation versus the displacement hypothesis. *Journal of Computer-Mediated Communication*, *12*(4), 1169–1182. https://doi.org/10.1111/j.1083-6101.2007.00368.x

Wagg, A. J., Callanan, M. M., & Hassett, A. (2019). Online social support group use by breastfeeding mothers: A content analysis. *Heliyon*, *5*(3), e01245. https://doi.org/10.1016/j.heliyon.2019.e01245

Walther, J. B. (2007). Selective self-presentation in computer-mediated communication: Hyperpersonal dimensions of technology, language, and cognition. *Computers in Human Behavior*, *23*(5), 2538–2557. https://doi.org/10.1016/j.chb.2006.05.002

Walther, J. B. & Burgoon, J. K. (1992). Relational communication in computer-mediated interaction. *Human Communication Research*, *19*(1), 50–88. https://doi.org/10.1111/j.1468-2958.1992.tb00295.x

Wang, Y. & Sandner, J. (2019). Like a "frog in a well"? An ethnographic study of Chinese rural women's social media practices through the WeChat platform. *Chinese Journal of Communication*, *12*(3), 324–339. https://doi.org/10.1080/17544750.2019.1583677

Wang, Y. C., Kraut, R. E., & Levine, J. M. (2015). Eliciting and receiving online support: Using computer-aided content analysis to examine the dynamics of online social support. *Journal of Medical Internet Research*, *17*(4), e99. https://doi.org/10.2196/jmir.3558

Weigold, A., Weigold, I. K., & Russell, E. J. (2013). Examination of the equivalence of self-report survey-based paper-and-pencil and internet data collection methods. *Psychological Methods*, *18*(1), 53. https://doi.org/10.1037/a0031607

Weinberg, J. D., Freese, J., & McElhattan, D. (2014). Comparing data characteristics and results of an online factorial survey between a population-based and a crowdsource-recruited sample. *Sociological Science*, *1*, 292–310. https://doi.org/10.15195/v1.a19

Wenz, A. (2021). Do distractions during web survey completion affect data quality? Findings from a laboratory experiment. *Social Science Computer Review*, *39*(1), 148–161. https://doi.org/10.1177/0894439319851503

Williams, T. A. & Shepherd, D. A. (2017). Mixed method social network analysis: Combining inductive concept development, content analysis, and secondary data for quantitative analysis. *Organizational Research Methods*, *20*(2), 268–298. https://doi.org/10.1177/1094428115610807

Woo, S. E., Keith, M., & Thornton, M. A. (2015). Amazon Mechanical Turk for industrial and organizational psychology: Advantages, challenges, and practical recommendations. *Industrial and Organizational Psychology*, *8*(2), 171. https://doi.org/10.1017/iop.2015.21

Wright, K. B. (2005). Researching Internet-based populations: Advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services. *Journal of Computer-Mediated Communication*, *10*(3), JCMC1034. https://doi.org/10.1111/j.1083-6101.2005.tb00259.x

Wright, K. B. (2016). Communication in health-related online social support groups/communities: A review of research on predictors of participation, applications of social support theory, and health outcomes. *Review of Communication Research*, 4, 65–87. https://doi.org/10.12840/issn.2255-4165.2016.04.01.010

Wright, K. B. (2017). Web-based survey methodology. In P. Liamputtong (ed.), *Handbook of Research Methods in Health Social Sciences* (pp. 1–14). Springer. https://doi.org/10.1007/978-981-10-2779-6_18-1

Wright, K., Fisher, C., Rising, C., Burke-Garcia, A., Afanaseva, D., & Cai, X. (2019). Partnering with mommy bloggers to disseminate breast cancer risk information: Social media intervention. *Journal of Medical Internet Research*, *21*(3), e12441. https://doi.org/10.2196/12441

Zhang, J., Calabrese, C., Ding, J., Liu, M., & Zhang, B. (2018). Advantages and challenges in using mobile apps for field experiments: A systematic review and a case study. *Mobile Media & Communication*, *6*(2), 179–196. https://doi.org/10.1177/2050157917725550

Zhou, H. & Fishbach, A. (2016). The pitfall of experimenting on the Web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, *111*(4), 493. https://doi.org/10.1037/pspa0000056

19 Archival Data

Jason Miller

Abstract

Social and behavioral researchers often draw on archival data – data collected by an entity other than the research team – to conduct scientific inquiry. Researchers typically seek to make claims about measured variables that extend beyond the measures themselves, such as interpreting a measure as representing an unobservable theoretical construct. Though researchers using archival data encounter many issues, this chapter focuses on two that have received less attention. The first concerns how researchers should justify the interpretations and uses they attach to archival measures. The second concerns how to justify generalizing findings. This chapter provides a framework to help researchers address these issues by drawing on contemporary validity theory in education and psychology as well as theory regarding causal mechanisms from philosophy and sociology. These concepts are illustrated using multiple examples from published studies.

Keywords: Archival Data, Validity, Mechanism, Generalizability

Introduction

The social and behavioral sciences are awash with archival data – data that were collected by an entity other than the research team. For example, the General Social Survey provides information about US adults' attitudes regarding multiple issues back to 1972 (General Social Survey, 2021); the American Community Survey provides detailed demographic information for the United States (Census Bureau, 2014); Open Secrets (www.opensecrets.org) compiles campaign finance data (Vegter et al., 2020); the Bureau of Labor Statistics publishes data on producer prices (Bureau of Labor Statistics, 2021b; Peltzman, 2000); and the Property Rights Alliance produces the International Property Rights Index (www.internationalpropertyrightsindex.org; Skowronski & Benton, 2018). In other instances, researchers may draw from archival sources (e.g., newspaper articles) and, using text-mining algorithms, convert qualitative data to quantitative data (e.g., construction of a measure of economic policy uncertainty; Baker et al., 2016). Likewise, data generated by companies in their normal course of operations can provide the raw ingredients for developing new measures (Scott, 2015, 2018; Winter et al., 2012).

Researchers using archival data usually desire to make claims that extend beyond the actual measures (Kane, 2013). One type of claim concerns whether archival measures can be interpreted as representing theoretical constructs (Bollen, 1989; Little, 2013). For example, in stating, "US economy has performed better when the president of the United States is a Democrat rather than a Republican," Blinder and Watson (2016, p. 1015) implicitly assume that archival gross domestic product (GDP) data can be interpreted as representing a broader construct of macroeconomic performance. While most would agree that a strong justification exists for interpreting GDP as macroeconomic performance, this illustrates a crucial point – the veracity of researchers' conclusions rests on the strength of the logical justifications for the interpretations and uses attached to archival measures (Kane, 2013). The second type of claim concerns how researchers generalize findings, especially when they have archival data from unique settings such as concrete manufacturing (Syverson, 2004) or truck transportation (Scott et al., 2021).

While the lack of control over data collection creates multiple challenges for researchers, this chapter brings attention to these two issues – making and justifying validity claims (Kane, 1992, 2001, 2013) and generalizing findings beyond the sample domain. Regarding the former, I adopt a unitary view of validity, which emphasizes whether the interpretation and use attached to observed measures are logically defensible (Kane, 2013; Messick, 1995), as opposed to treatments that emphasize different forms of validity (e.g., content validity, convergent validity, criterion validity, and discriminant validity; Bollen, 1989; Foster & Cone, 1995). The unitary view is the preferred conceptualization in education and psychology, fields where concerns about validity have been especially salient given the high-stakes nature of standardized testing and psychological diagnoses (Kane, 2013; Messick, 1995). The unitary perspective also reduces the ability of researchers – the present author included – to "cherry pick" forms of validity (Miller et al., 2021c). Regarding generalization, this chapter emphasizes the central role theoretical mechanisms play in supporting claims that effects are likely to occur in other settings (Astbury & Leeuw, 2010; Mahoney, 2001). Importantly, it does not address whether researchers can justify statistical conclusions or causal claims (Shadish et al., 2002) made based on analyses of archival data. The reason is justifications of this sort are based on research design characteristics (Angrist & Pischke, 2010) and the appropriateness of a specific statistical model for answering the question of interest (Cudeck & Henly, 1991), topics beyond the scope of this chapter.

To provide concrete examples of abstract ideas, this chapter uses Baker et al.'s (2016) development of a measure of economic policy uncertainty (EPU) as a running example due to the impact this paper has in economics. Their article represents a unique application of archival data – they construct their main EPU measure using frequency counts of articles in ten major US newspapers based on articles containing combinations of key terms. As a result, they assemble a monthly EPU measure for the United States back to 1985. This said, the arguments apply regardless of data structure. Archival measures can be time-series data for a single variable (Enders, 2015), such as personal income in the United States (Bureau of Economic Analysis, 2021), cross-sectional data collected across multiple subjects at a single point in time, such as the dependence of a firm's supply base for a given year (Schwieterman et al., 2020), or panel data collected across time for multiple subjects (Singer & Willett, 2003). However, before concerns about validity enter the discussion, researchers much obtain data from archival sources – the issue I first address.

Obtaining Archival Data: Some Personal Experiences

A statement I frequently hear from individuals is that they think obtaining archival data is “far easier than collecting your own data.” When I hear such statements, I immediately know the speaker has little – and more likely no – experience working with archival data. As such, and to hopefully save you some time (and heartache), I want to offer a few tips for researchers just beginning their journey.

Regarding finding archival sources, I have found three strategies to be especially useful. The first is to read broadly on your topic area in both your discipline’s most respected journals as well as the respected journals in other disciplines. Keep a list of data sources that may be of interest and check whether you have access to these through your employer. My experience is that researchers who study your topic through a different disciplinary lens are more likely to utilize data sources that you have less familiarity with, consistent with Granovetter’s (1973) strength of weak ties thesis. The second is to search through data repositories at major government agencies (e.g., the Bureau of Labor Statistics, Census Bureau, Bureau of Economic Analysis, etc.). The third is to conduct searches on statistical data compilation websites (e.g., Statista) to identify relevant data series, and then track down the original sources to obtain more information. Especially when working with data from non-random samples, I urge researchers to utilize the framework presented by Brave et al. (2021) to evaluate whether the data fit their needs.

Regarding obtaining archival data, especially when data are from private entities, my recommendation is to reach out to the individual or organization, explain that you are an academic researcher, and ask if they would be willing to share data for research purposes. Reaching out in this manner led to me obtaining archival truck driver turnover data from the American Trucking Associations that my colleagues and I have since used in two studies (Miller et al., 2020, 2021a) to answer research questions that could not be answered with cross-sectional primary data, which had been the *modus operandi* for studying truck driver turnover. The key thing to remember is that you need to be able to explain how you can offer value to the entity that generates the data.

Another recommendation about obtaining archival data is that, depending on your research design, sometimes brute-force data entry, while monotonous to say the least, is the way to go. For example, prior to a co-author with a strong computer science background automating data collection efforts, I spent hundreds of hours in the last year of my PhD program and the first three years as an assistant professor manually collecting longitudinal safety compliance data from the Department of Transportation for hundreds of trucking companies across four different sampling frames. While I will be the first to admit that this likely was not an optimal use of my time, the net result has been ten high-level publications. As such, do not be afraid to get your fingers dirty in the pursuit of archival data.

Validity of Archival Data

Messick (1995, p. 741) defines validity as, "an overall evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other modes of assessment." In this context, "test scores" and "other modes of assessment" refer to measures obtained directly from archival sources (e.g., producer prices from the Bureau of Labor Statistics). Examining this definition, a few features are worth noting. First, validity concerns the strength of justification for the interpretation attached to a variable (e.g., Measure *X*), not Measure *X* itself (Kane, 1992). This implies that one interpretation attached to Measure *X* may be deemed reasonably valid, whereas another interpretation is not. For example, a strong justification can be made to interpret firms' patent counts as a measure of firms' innovations, but this justification cannot be made to interpret patent counts as representing firms' technological capabilities (Ketchen et al., 2013). This issue is especially salient with archival data because these data are often collected by actors who have no intention of measuring theoretical constructs; this can drive researchers to attach questionable interpretations to archival data. As an example, researchers trying to test predictions of transaction cost economics (Williamson, 2005) have argued that measures such as firms' advertising intensity can be interpreted as representing asset-specific investments – a theoretical construct central to transaction cost economics – even though such an interpretation can be challenged on conceptual grounds (Ketchen et al., 2013).

Second, researchers developing the justifications regarding a specific interpretation to Measure *X* are not limited to offering statistical evidence (Messick, 1995). Rather, justifications of why a given interpretation can be attached to Measure *X* are often grounded in existing theory (Cizek, 2012). For example, Basu (2019) draws on microeconomic theory to justify why a firm's markup price over marginal cost (Measure *X*) can be interpreted as a measure of market power (Theoretical Construct X'). Similarly, Scott & Nyaga (2019) explain why a subset of truck drivers' hours-of-service violations (Measures $X_1 - X_n$) can be utilized to represent the more abstract theoretical construct of intentional rule violations (Theoretical Construct X') by detailing the underlying causal process of how these violations occur and explaining how this process aligns with motive–opportunity–choice theory (McKendall & Wagner, 1997). Another type of evidence is if practitioners (e.g., stock analysts) utilize the Measure *X* analogous to the meaning of Theoretical Construct X' – also known as "vetted by the market" evidence (Baker et al., 2016).

Third, an interpretation is not attached to a measure in isolation; rather, researchers assign an interpretation to Measure *X* for a specific use (Kane, 2001). Many times, the use involves generalizing the given score to some broader domain (e.g., arguing why Measure *X* can be utilized to represent Theoretical Construct X' in theory testing; Kane, 1992). So long as theory testing studies do not disclose individual subjects' data, the hurdle for using Measure *X* for theory testing is rather low. In contrast, if data are to be released, especially data that rank subjects, greater justification is required because, as law-school rankings have demonstrated

(Espeland & Sauder, 2007; Sauder & Espeland, 2009), such data can have major practical consequences for the entities being ranked. The possibility of use beyond academic research (e.g., publishing a new measure online) is more pressing with archival data because primary data are often protected from disclosure through institutional review board protocols; archival data are less likely to be subject to these limits.

Fourth, interpretations are practical arguments that should be judged on the grounds of argument clarity, argument coherence, and the plausibility of assumptions underlying the argument (Cizek, 2012; Kane, 1992). Messick (1995, p. 742) states, "validation combines scientific inquiry with rational argument to justify (or nullify) score interpretation and use." As such, validity claims regarding archival data cannot be evaluated on absolute terms – they rest on a continuum from poorly to well supported (Cizek, 2012; Kane, 2013).

With these points in mind, the next subsections provide more detail on the specific facets of validity that authors should take into consideration when utilizing archival data. As detailed by Cook and Beckman (2006), Downing (2003), and Messick (1995), these five facets are (i) content evidence, (ii) response process, (iii) internal structure, (iv) relations to other variables, and (v) consequences of use. The first four of these are discussed in the subsections below. Researchers' evidential strength in each facet need not be equally compelling – my experience suggests there are often trade-offs across facets. However, as noted by Messick (1995, p. 744), "What *is* required is a compelling argument that the available evidence justifies the test [archival] interpretation and use, even though some pertinent evidence had to be forgone," [emphasis original]. When possible, examples from different disciplines are given to illustrate the principles but, as the reader will surely notice, my greater knowledge of economics and business research will skew the examples toward this domain.

Content Evidence

The content evidence facet of validity concerns two aspects regarding how well Measure X serves to represent Theoretical Construct X'. The first aspect is the extent Measure X fully taps the meaning of Construct X'. An issue researchers often encounter is that archival measures do not fully tap a theoretical construct's domain. A setting where this occurs is studying the theoretical construct of *Power* as it pertains to buyer–supplier relationships. Frazier (1983, p. 158) defined *Power* as, "the ability of one channel member to influence decision variables of another channel member, a potential for influence on another firm's beliefs and behavior." Given Emerson's (1962) argument that the power Actor A has on Actor B is equal to Actor B's dependence upon Actor A, researchers testing theories regarding *Power* often seek to measure *Dependence*. Efforts to do this using archival sources have been challenging, and this includes my own work (Schwieterman et al., 2020). For example, a highly cited study by Casciaro and Piskorski (2005, pp. 183–184) measures industry *M*'s dependence on industry *N* by first summing the percentage of industry *M*'s total sales that industry *N* buys plus the percentage of industry *M's* purchases that industry *N* supplies; they then multiply this summed value by industry *N*'s four-firm concentration ratio. As this measure cannot capture industry *M's* outside options, both regarding customers and supply sources, a key facet of *Dependence* and, consequently, *Power*, is uncaptured because outside options have an important impact on one firm's ability to influence another firm (Heide & John, 1988). However, this deficiency is counterbalanced by the fact that the authors have data measured in a consistent way over many years from a highly reliable government source (Bureau of Economic Analysis) – that would qualify as a strength of the response process facet of validity.

Another aspect of content evidence concerns the extent that Measure X contains unwanted variance that is unrelated to Theoretical Construct X'. One form of unwanted variation, especially prevalent with archival data, occurs when archival measures represent broader aggregates than the theoretical construct the researcher intends on representing. For example, in Miller et al. (2018), my colleagues and I were interested in testing how trucking companies' use of independent contractors affected company-level scores on the physical condition of their equipment – this falls within the general theoretical domain of the construct *Equipment Maintenance* (McKone & Weiss, 1998). To do this, we utilized archival data from the Safety Measurement System Program – a measure called the Vehicle Maintenance BASIC Score. The unwanted variance issue was that the Vehicle Maintenance BASIC Score also included violations for shipments not being properly secured. As issues pertaining to load securement are outside the domain of the theoretical construct *Equipment Maintenance*, this raises the concern of interpreting the Vehicle Maintenance BASIC Score as representing *Equipment Maintenance*. Fortunately, discussion with regulators who had access to the underlying data confirmed that these load securement violations represented a small percentage of the total violations within this category; this suggested that the variance in the Vehicle Maintenance BASIC Score was minimally affected by load-securement violations.

Baker et al. (2016) demonstrate additional ways in which researchers can provide evidence regarding this facet of validity. As their study relies on an automated text-based search of newspaper articles, there is the concern that the algorithm is selecting articles that have little to do with EPU. To address this, the authors created a large audit study that compares the performance of machine versus human coders. As shown in their manuscript, the two approaches agree very closely and provide evidence that their automated approach is appropriately capturing the content they wish to measure.

To conclude the discussion of content evidence, I would like to offer the following two points of advice for authors working with archival data:

- Seek clarification regarding the underlying components that serve as inputs into aggregate scores. If possible, obtain information regarding the weight each component contributes to the aggregate to ensure that an unacceptable degree of unwanted variation is not being produced.
- Be forthcoming about how well the observed measures can capture the full domain of a theoretical construct. In some instances, it may be necessary to develop

a validity claim that an observed archival measure represents a narrower theoretical construct for which a stronger validity claim can be made.

Response Process

The response process facet of validity concerns whether the archival data were generated through a stable, repeatable process that reduces the likelihood that idiosyncrasies contaminate the measures (Cook & Beckman, 2006; Miller et al., 2021c). One key issue is whether archival data are self-reported. When data are self-reports, it is important to examine if the self-reporter has an incentive to misreport information (e.g., about employee accident rates) and, if so, whether there are checks in place to prevent misreporting. Illustrating this concern, Forbes et al. (2015) demonstrate that, prior to automated computer reporting of aircraft arrival times, airlines tended to misreport their aircraft arrival times to inflate their on-time arrival rates. The airlines had an incentive to do this because airlines' company-wide arrival rates are publicly reported by the Department of Transportation. Illustrating the principle of checks for misreporting is the fact that companies listed on US stock exchanges must have their financial statements audited by independent firms, with top executives facing serious repercussions for fraudulent reporting. Likewise, firms are legally required to complete documents (e.g., the Economic Census – conducted by the Census Bureau) and face legal ramifications if inaccurate information is intentionally reported (Ali et al., 2008).

Archival data that are not self-reports can also exhibit response process concerns. For example, Jin and Leslie (2003) document that a change in Los Angeles' requirements that restaurants disclose their hygiene scores to customers resulted in inspectors changing their behavior – inspectors became more likely to score restaurants just above the threshold necessary to receive an "A" rating. Likewise, when archival data represent ratings that are about abstract concepts, such as how well a firm executes quality management processes, it is important for researchers to explain whether a standardized process exists by which scores are generated and whether there is triangulation across multiple raters. For example, Miller and Parast (2019) describe that each of the seven quality subdimensions scored when firms apply for the Malcolm Baldrige National Quality Award are rated independently by six to ten quality experts; the median score is utilized for each subdimension. The existence of multiple independent raters assuages concerns that methods (e.g., which person does the rating) are what shape the variation in Measure *X*.

Another response process issue that affects some archival data is that some subjects may be observed infrequently or more frequently than others. An example of the former issue occurs with archival studies regarding restaurant hygiene (Jin & Leslie, 2003). Since there is strong evidence that inspectors are idiosyncratic (Macher et al., 2011) and that factors such as the timing of an inspection within a regulator's workday affect whether violations are detected (Ibanez & Toffel, 2020), researchers should recognize that data for firms with a limited number of inspections are likely to be noisy. As such, when working with archival data for which scores were generated from a limited number of records, researchers are well served to provide additional evidence that measurement error is not unduly affecting findings (Miller & Saldanha, 2018).

An example of the latter issue concerns school-level average standardized test scores for a particular grade. As explained by Kane and Staiger (2002), caution is warranted in drawing strong conclusions about schools displaying different levels of standardized test performance when some of the schools in question have a small number of pupils completing the exams. The challenge is that small schools' average scores are measured less reliably than larger schools' scores due to random errors canceling out in larger schools. This can pose special challenges if researchers seek to study changes in scores over time, as the estimated changes may have little meaning. One strategy to address this issue is to weigh residuals based on the number of underlying records that compose an aggregate (e.g., using the square root of the number of students in a grade), as this places greater weight on records that should be measured with a higher degree of reliability.

The entity generating the data may also rely heavily on imputation approaches to address missing data. Imputation here refers to the data-generating entity filling in missing values using some systematic technique. For example, the Census Bureau and Bureau of Transportation Statistics rely extensively on imputation to address missing data issues in the Commodity Flow Survey (Census Bureau, 2020). Similarly, a different division of the Census Bureau relies on a new imputation technique to calculate state-level monthly retail sales data (Census Bureau, 2021b).

Response process concerns manifest in Baker et al.'s (2016) study due to newspapers having different political slants that may affect how they cover certain topics. To address this possibility, the authors categorize their ten papers based on the extent they slant Republican versus Democratic using an index of media slant from Gentzkow and Shapiro (2010). They then recalculate their EPU measure for these two subsets of newspapers and report a correlation of 0.92, suggesting that political slant of newspapers does not represent a response process element that is driving their results.

As with the prior section, I conclude this subsection with a few points to keep in mind regarding the response process facet of validity as it pertains to archival data:

- While archival data are often treated as distinct from surveys, researchers should remember that archival data often start life as surveys. For example, most economic data that are collected by the Census Bureau and Bureau of Labor Statistics come from surveys (Horowitz & Planting, 2009), even though researchers would describe these data as archival. Thus, issues associated with surveys, such as informants interpreting questions differently – a key response process concern in survey research (Downing, 2003) – can also apply with archival data (Census Bureau, 2021a).
- Obtain as detailed information as possible regarding the process through which the archival data were generated. Government agencies usually have publicly available extensive documentation regarding the processes they use (e.g., Bureau of Labor Statistics, 2021a). If archival data come from private sources, ask (within

confidentiality bounds) to know as much about the data as possible. It is also useful to gain commitment from private sources that they will continue to engage throughout a project's review process, such as providing additional information to address reviewers' concerns.

- Understand the extent to which procedures are in place to ensure a consistent data-generating process with controls for unusual responses. For example, researchers should know if there is a standardized scoring rubric, if multiple raters are utilized, and if raters have received similar training (Downing, 2003). The existence of standardized rubrics and multiple raters is especially important when the scores are for abstract concepts (e.g., quality management competence), as opposed to concrete concepts (e.g., dollars of sales). Furthermore, researchers should seek evidence about whether procedures exist to flag unusual observations to rectify these cases.

Internal Structure

The internal structure facet of validity applies to settings where researchers have multiple measures (Measures $X_1 - X_n$) that are argued to be different manifestations of the same theoretical construct (i.e., reflective measures; Bollen, 2002). Researchers would thus expect these measures to be correlated with one another. The internal facet structure of validity thus refers to the psychometric characteristics of Measures $X_1 - X_n$ (Downing, 2003). These psychometric characteristics include the magnitude of factor loadings (Wirth & Edwards, 2007) and whether the measures display similar psychometric properties across groups (Downing, 2003). Researchers have shown great creativity in merging multiple data sets using techniques, and I refer readers to Bauer and Hussong (2009) for a fascinating application. As outstanding treatments exist for factor analysis (Browne, 2001) and item response theory (Wirth & Edwards, 2007), I do not delve into these techniques. Instead, I want to focus on a few more subtle issues that I have encountered when applying psychometric techniques to archival measures.

- Relative to surveys using multi-item scales to measure abstract constructs, archival measures may concern a very narrow sampling domain; this can result in measures having very high correlations. For example, in Muir et al. (2019), three separate archival measures capturing trucking companies' service specialization displayed an average correlation ≥ 0.95. When a common factor model is over-identified, such high correlations can result in estimated measurement models fitting well, based on examining residuals, but the maximum likelihood discrepancy function (and consequently any fit indices that incorporate said discrepancy function) suggesting severe misfit (Browne et al., 2002). Without diving into the mathematics (see MacCallum et al., 2002), this situation can surprise researchers who are not aware of this phenomenon.
- Be careful about directly combining different archival measures that have dramatically different degrees of variance. As explained by Cudeck (1985), challenges can be encountered when fitting factor analysis models where the

measures rest on very different scales. One solution is to place the observed scores on the same scales, but caution is warranted in that different transformation approaches can affect results. For example, transforming Measure X_1 and Measure X_2 such that they both rest on a 0–1 sample space will help ensure the standard deviations are the same while allowing the measures to still have different variances. In contrast, normalizing X_1 and X_2 will result in each having the same variance of 1. Thus, researchers must carefully consider how transformations affect their data's distributions.

- Recent advances in multigroup techniques, especially the alignment method (Asparouhov & Muthén, 2014; Muthén & Asparouhov, 2014, 2018) have increased researchers' abilities to test for differential item functioning across groups. To the extent that researchers' data structure permits such analysis, this technique may be useful to assuage concerns that results are being unduly affected by archival measures operating differently across groups.
- There is increased interest in applying factor analytic models to multivariate time-series data (Asparouhov & Muthén, 2020; Asparouhov et al., 2018; Hamaker et al., 2018). In addition to providing unique information about the internal consistency of multiple time series, these applications may present unique opportunities for researchers to push the bounds of existing theory with archival data.

Relations to Other Variables

The relations to other variables facet of validity concerns whether Measure X is correlated with one or more other measures in a manner consistent with extant theory. Existing theory informs this facet of validity because theory usually suggests some pattern of relations between theoretical constructs. Consequently, if researchers wish to claim that they have observed measures that can be interpreted as representing these theoretical constructs, a relevant piece of evidence is whether their measures show the same pattern of relations as predicted by theory. For example, Baker et al. (2016) document that their newspaper-based measure of EPU correlates 0.73 with 30-day options-implied volatility of the S&P 500 volatility index and 0.54 with frequency counts of the word "uncertain" in the Federal Open Market Committee's Beige Book. The authors further document that EPU negatively affects employment growth and investment, especially for firms in industries that are more exposed to government purchases. The authors then demonstrate, using vector autoregressions, that a positive EPU shock negatively affects employment and industrial production. As these findings are consistent with economic theory, they provide further evidence for the validity of Baker et al.'s (2016) measure.

Another example of how researchers have leveraged this facet of validity is to evaluate which archival source can be better used to measure a given theoretical construct. An excellent example of this practice is Ali et al. (2008), who re-examine the use of industry concentration measures calculated from only publicly traded companies (Compustat) relative to using industry concentration measures from the

Economic Census – which include all public and private firms. Ali et al. (2008) suggest that a stronger validity claim can be made for the Economic Census concentration measures better representing the theoretical construct *Industry Concentration* because the Economic Census concentration measures correlate with other variables in a manner more consistent with theory vis-à-vis those from Compustat.

Readers are likely to have two questions at this stage. First, how are we to decide which measures we should have more confidence in when making these evaluations? Second, do relations between variables better belong in the results section? Beginning with the former, my answer is that researchers usually have more confidence in the validity claims underlying some measures vis-à-vis others. For example, the data needed to measure one theoretical construct may be more concrete, and this reduces response process concerns regarding reporting. For example, the employment data utilized by Baker et al. (2016) to study the consequences of their EPU measure is concrete and available for administrative records; this suggests employment data are less affected by measurement issues. In instances where this is not feasible, my argument rests on principles from inference to the best explanation (Lipton, 2004). Imagine that a researcher has Measures *X* and *Y* where validity claims have been advanced that they represent Theoretical Constructs *X′* and *Y′*, respectively. Existing theory predicts *X′* and *Y′* are positively correlated. After collecting data, our research team finds that *X* and *Y* are positively correlated. As this evidence is consistent with our validity claims, it can serve as an additional input into the researcher's argument-based validation approach (Kane, 2013), provided there is not some alternative compelling reason why this positive correlation exists (e.g., common method concerns). The skeptic must advance an equally compelling alternative explanation as to why these relations exist, which does not rest on these measures representing their theoretical constructs, as inference to the best explanation is always comparative (Lipton, 2004).

Turning now to the second question, I will admit that relations between variables often do take the research team into the results section. However, there is no logical reason why evidence from later in a manuscript cannot be utilized to reinforce our confidence that we are indeed measuring what we are arguing to measure. For example, imagine theory suggests that Theoretical Construct *X′* will have a positive average relationship with Theoretical Construct *Y′*, but this relationship will be reduced when Theoretical Construct *M′* is high (implying a negative two-way interaction; Aiken & West, 1991). A research team draws on one or more archival sources to obtain Measures *X*, *M*, and *Y*. Holding constant other relevant covariates that reside theoretically upstream from *X* and *M*, and are also uniquely partially correlated with *Y*, our researchers find evidence that there is a negative two-way interaction between *X* and *M*. This seems like a mighty strange set of coincidences, to paraphrase Meehl (1990), to observe if *X*, *M*, and *Y* are not capturing the theoretical constructs they are being argued to represent. As such, why would a researcher not be allowed to include such information to support the validity claims?

In concluding this subsection, I would like to offer a few additional thoughts regarding how researchers can best leverage the relations to other variables facet of validity to make claims:

- Focus attention on the magnitude of relations between observed measures, not simply whether statistically significant relationships exist. As explained by Meehl (1990), theory often points toward some theoretical constructs showing different magnitude relations. Finding evidence of such effects can offer particularly compelling evidence that observed measures can be interpreted as representing theoretical constructs – it is difficult to conceive of alternative explanations that could have brought about the observed relationships. I have frequently exploited this principle in my own research (Miller & Saldanha, 2016; Miller et al., 2018). This strategy can also be utilized to help assuage concerns about endogeneity due to omitted right-hand-side variables (Bloom et al., 2012; Miller et al., 2022).
- Relations between measures obtained from different sources may provide stronger evidence to the extent that this can assuage concerns that common method effects are driving observed relationships. However, this must be balanced by concerns about different archival sources measuring data at different levels of aggregation.

A Strategy for Validating the Interpretation & Use of Measures from Archival Data

Kane (2001) presents a concise, four-step outline regarding how researchers can validate the proposed interpretation and use of observed measures. The first step is for researchers to specify the proposed interpretation and use of their measure(s). The second is to compile logical and statistical evidence that supports this interpretation and use, paying special attention toward underlying assumptions that are the most problematic. The third step is to collect and evaluate any additional evidence concerning the most problematic assumptions identified in the second step. The fourth step is to iterate through the first three steps to refine the proposed interpretation and use based on available logical and empirical evidence.

As this process is abstract, I want to share a personal example to illustrate the process. In Miller et al. (2021b), my co-authors and I examined how changes in prices for spot market truckload shipments affected the contract price of truckload shipments. Thus, we are interested in testing how Theoretical Construct X' (*Truckload Spot Prices*) affects Theoretical Construct Y' (*Truckload Contract Prices*). Measure X is monthly dry van spot market truckload prices from DAT Freight & Analytics (DAT Freight & Analytics, 2021), whereas Measure Y is the Bureau of Labor Statistics producer price index (PPI) for general freight, long-distance, truckload firms (Bureau of Labor Statistics, 2021c). For brevity, I focus only on the evidence we utilized to support interpreting the PPI as representing *Truckload Contract Prices*. A key issue is that the PPI does not separate shipments priced on a spot basis versus those priced according to long-term contracts. This raised the concern that the statistical relationship we identified were the result of PPIs also capturing spot price movements.

To evaluate this concern, we turned to several pieces of evidence. First, the month-over-month percent changes in the PPI data were far less than the month-over-month percent changes for DAT's data, that were only spot prices. This was reassuring since prior research that had access to company-level contract and spot prices showed contract prices were much more stable (Bai, 2018). This suggested, at minimum, that the PPI data were capturing contract prices to a much greater degree than spot prices. Second, and even more compelling, we found distributed lagged effects (Almon, 1965) suggesting a change in DAT's spot price data in month *t* affected the PPI over a course of five months (as opposed to only affecting the PPI in the current month). This finding was aligned with existing theory that contract prices take time to incorporate supply and demand dynamics conveyed by spot prices. Furthermore, if the effect stemmed from the PPI's inclusion of spot prices, this effect should be fully captured in the present month (i.e., there would not be a distributed lagged pattern). Third, we found evidence that this effect became more pronounced over time following the start of a specific regulatory change. However, the composition of spot versus contract freight did not dramatically shift at this date – this eliminates the possibility that the composition of the PPI rapidly shifted after this regulatory intervention. Taken together, these independent pieces of evidence allowed us to adequately justify that the PPI could be interpreted as representing *Truckload Contract Prices*.

Generalizing Findings from Archival Data Studies

Once researchers have completed the arduous process of validating the interpretations they have assigned to their measures and estimated their statistical models, they face the challenge of generalizing their results beyond their sample. This could entail generalizing findings beyond the time frame covered by the study (e.g., assuming results from the 1980s hold today), to a broader population than was studied (e.g., assuming results from publicly traded companies hold for private companies), or across countries (e.g., assuming productivity findings from European firms hold in American firms). This raises the natural question: What evidence should researchers draw on to defend their generalizations? This issue is especially important with archival data because researchers (i) may only have access to a very limited number of subjects (e.g., data from a single firm; Scott, 2015, 2019) or data that are many years old (Braguinsky et al., 2015) and (ii) cannot collect similar data for more subjects or recent times.

My answer to this question has both a methodological and theoretical slant. Concerning the former, one question researchers must ask is: What is the coverage of the archival data that we are using? Brave et al. (2021) provide an example in explaining that data from the scheduling software provider Homebase showed a much more dramatic drop in employment at the start of the COVID-19 pandemic than official figures from the Bureau of Labor Statistics because Homebase's sample is skewed toward smaller firms in hard-hit industries (e.g., restaurants). A similar concern that pertains to archival data collected via surveys is the response rates for said surveys. To the extent that non-response may be associated with observed characteristics (e.g., smaller establishments are less likely to respond), researchers should be cautious in generalizing their findings to establishments that have the characteristics associated with the non-response.

The most important factor affecting generalization is the confidence that researchers have that the underlying mechanisms theorized between the theoretical constructs will hold in other settings (Hedström & Ylikoski, 2010). As with devising validity claims, researchers must be able to present cogent arguments as to why the mechanisms undergirding their theories apply to the domain of generalization (Steel, 2004). This is made more challenging because mechanisms are unobservable (Astbury & Leeuw, 2010; Bunge, 2004) and, hence, cannot be measured with quantitative archival data. The crux centers on whether contextual characteristics exist in the domain of generalization that could deactivate the mechanism(s) theorized to operate in the present setting (Falleti & Lynch, 2009; Pawson & Manzano-Santaella, 2012). If contextual factors in the domain of generalization are known to differ substantially from the setting at hand, and there is good reason to believe these differences increase the likelihood that the mechanisms will not operate, then generalization is unwarranted. Conversely, if researchers can present a strong case that contextual factors in the domain of generalization support the mechanism(s) activation, greater confidence can exist for generalizing findings.

The theoretical emphasis on mechanisms as it pertains to generalizing findings is important because it may allow researchers to generalize findings that may, at first glance, seem highly idiosyncratic. Take, for example, Braguinsky et al.'s (2015) study of the performance consequences of acquisitions in the Japanese cotton spinning industry from 1896 through 1920. Seeing this unique historical setting, readers may be skeptical that the authors' findings apply to today's business environment. To assuage these concerns, the authors highlight how the economy of Japan at the turn of the twentieth century was very akin to today's Western capitalistic economies, especially emphasizing how the ownership and control structures mirror modern firms. They then explain that the similarity of contextual factors should allow for their management diffusion mechanism to operate in today's business context. Similarly, Syverson's (2004) use of the ready-mix concrete sector to study how *Demand Density* affects the probability distribution of firms' *Productivity* could raise concerns about generality, given the unique features of this manufacturing setting (e.g., competition is geographically isolated because it is economically infeasible to transport concrete long distances). However, such concerns are assuaged by the author, clearly articulating the underlying mechanisms and providing evidence that these mechanisms are highly general.

These examples illustrate two key principles that underlie effective generalization with archival data. These are:

- Clearly articulate the mechanism(s) postulated to bring about the relationships between theoretical constructs. Unfortunately, as noted by Sutton and Staw (1995), researchers are often reluctant to do this because they fear reviewers will call into question their empirical models because unobservable mechanisms are not included (Astbury & Leeuw, 2010). Clearly articulated mechanisms both strengthen theory and allow other research teams to identify contexts where further pursuit of the theory can be most fruitfully conducted (Nyrup, 2015).

- Identify the contextual features that can activate or suppress mechanisms. Recognizing the conditions necessary to activate a mechanism strengthens theory by suggesting boundary conditions for the proposed relationships (Goldsby et al., 2013). In particular, the absence of a necessary contextual factor may suppress a mechanism and, consequently, result in a null relationship between theoretical constructs (Pawson & Manzano-Santaella, 2012). Applied to discussions about generalizing findings, this suggests that researchers should avoid generalizing findings to domains where necessary contextual factors to activate a mechanism are absent.

Conclusion

Researchers in the social and behavioral sciences will continue to rely heavily on archival sources to test their theories and, in doing so, push the boundaries of knowledge. Doing this, however, requires researchers to clearly articulate the interpretations they attach to archival data and provide convincing evidence that these data can represent the theoretical constructs that are at the heart of their theories. Researchers, likewise, must be able to cogently argue the domains to which their results generalize. The aim of this chapter has been to provide new and experienced users of archival data with a different perspective regarding how to best achieve these goals.

References

Aiken, L. S. & West, S. G. (1991). *Multiple Regression: Testing and Interpreting Interactions*. SAGE Publications.

Ali, A., Klasa, S., & Yeung, E. (2008). The limitations of industry concentration measures constructed with Compustat data: Implications for finance research. *Review of Financial Studies*, *22*(10), 3839–3871.

Almon, S. (1965). The distributed lag between capital appropriations and expenditures. *Econometrica*, *33*(1), 178–196.

Angrist, J. D. & Pischke, J. S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, *24*(2), 3–30.

Asparouhov, T. & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495–508.

Asparouhov, T. & Muthén, B. (2020). Comparison of models for the analysis of intensive longitudinal data. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(2), 275–297.

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(3), 359–388.

Astbury, B. & Leeuw, F. L. (2010). Unpacking black boxes: Mechanisms and theory building in evaluation. *American Journal of Evaluation*, *31*(3), 363–381.

Bai, X. (2018). Forecasting short term trucking rates. Unpublished Master's Thesis, Massachusetts Institute of Technology. Available at: https://dspace.mit.edu/handle/1721.1/117796.

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *Quarterly Journal of Economics*, *131*(4), 1593–1636.

Basu, S. (2019). Are price-cost markups rising in the United States? A discussion of the evidence. *Journal of Economic Perspectives*, *33*(3), 3–22.

Bauer, D. J. & Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods*, *14*(2), 101–125.

Blinder, A. S. & Watson, M. W. (2016). Presidents and the US economy: An econometric exploration. *American Economic Review*, *106*(4), 1015–45.

Bloom, N., Sadun, R., & Van Reenen, J. (2012). Americans do IT better: US multinationals and the productivity miracle. *American Economic Review*, *102*(1), 167–201.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. John Wiley & Sons.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, *53*(1), 605–634.

Braguinsky, S., Ohyama, A., Okazaki, T., & Syverson, C. (2015). Acquisitions, productivity, and profitability: Evidence from the Japanese cotton spinning industry. *American Economic Review*, *105*(7), 2086–2119.

Brave, S. A., Butters, R. A., & Fogarty, M. (2021). The perils of working with big data and a SMALL checklist you can use to recognize them. *Business Horizons*, *65*(4), 481–492. https://doi.org/10.1016/j.bushor.2021.06.004

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*(1), 111–150.

Browne, M. W., MacCallum, R. C., Kim, C. T., Andersen, B. L., & Glaser, R. (2002). When fit indices and residuals are incompatible. *Psychological Methods*, *7*(4), 403–421.

Bunge, M. (2004). How does it work? The search for explanatory mechanisms. *Philosophy of the Social Sciences*, *34*(2), 182–210.

Bureau of Economic Analysis (2021). Personal income. Available at: https://fred.stlouisfed.org/series/PI.

Bureau of Labor Statistics (2021a). Handbook of methods. Available at: www.bls.gov/opub/hom/home.htm.

Bureau of Labor Statistics (2021b). Producer price indexes. Available at: www.bls.gov/pPI/.

Bureau of Labor Statistics (2021c). Producer price index by industry: General freight trucking, long-distance truckload (PCU484121484121). Available at: https://fred.stlouisfed.org/series/PCU484121484121.

Casciaro, T. & Piskorski, M. J. (2005). Power imbalance, mutual dependence, and constraint absorption: A closer look at resource dependence theory. *Administrative Science Quarterly*, *50*(2), 167–199.

Census Bureau (2014). American community survey design and methodology (January 2014). Available at: www2.census.gov/programs-surveys/acs/methodology/design_and_methodology/acs_design_methodology_report_2014.pdf.

Census Bureau (2020). 2017 Commodity flow survey methodology. Available at: www2.census.gov/programs-surveys/cfs/technical-documentation/methodology/2017cfsmethodology.pdf#:~:text=%20%20%20Title%20%20%202017%

20Commodity,Created%20Date%20%20%201%2F22%2F2021%2012%3A21%3A48%20PM%20.

Census Bureau (2021a). Economic census, technical documentation, methodology, nonsampling error. Available at: www.census.gov/programs-surveys/economic-census/technical-documentation/methodology.html#nonsampling-error.

Census Bureau (2021b). Monthly state retail sales technical documentation. Available at: www.census.gov/retail/mrts/www/statedata/msrs_technical_documentation.pdf.

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods*, *17*(1), 31–43.

Cook, D. A. & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *American Journal of Medicine*, *119*(2), 166.e7–166.e16.

Cudeck, R. (1985). A structural comparison of conventional and adaptive versions of the ASVAB. *Multivariate Behavioral Research*, *20*(3), 305–322.

Cudeck, R. & Henly, S. J. (1991). Model selection in covariance structures analysis and the" problem" of sample size: A clarification. *Psychological Bulletin*, *109*(3), 512–519.

DAT Freight & Analytics (2021). National van rates. Available at: www.dat.com/industry-trends/trendlines/van/national-rates.

Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education*, *37*(9), 830–837.

Emerson, R. M. (1962). Power-dependence relations. *American Sociological Review*, *27*(1), 31–41.

Enders, W. (2015). *Applied Econometric Time Series*, 4th ed. John Wiley & Sons.

Espeland, W. N. & Sauder, M. (2007). Rankings and reactivity: How public measures recreate social worlds. *American Journal of Sociology*, *113*(1), 1–40.

Falleti, T. G. & Lynch, J. F. (2009). Context and causal mechanisms in political analysis. *Comparative Political Studies*, *42*(9), 1143–1166.

Forbes, S. J., Lederman, M., & Tombe, T. (2015). Quality disclosure programs and internal organizational practices: Evidence from airline flight delays. *American Economic Journal: Microeconomics*, *7*(2), 1–26. https://www.aeaweb.org/articles?id=10.1257/mic.20130164.

Foster, S. L. & Cone, J. D. (1995). Validity issues in clinical assessment. *Psychological Assessment*, 7(3), 248–260.

Frazier, G. L. (1983). On the measurement of interfirm power in channels of distribution. *Journal of Marketing Research*, 20(2), 158–166.

General Social Survey (2021). About the GSS. Available at: https://gss.norc.org/About-The-GSS.

Gentzkow, M. & Shapiro, J. M. (2010). What drives media slant? Evidence from US daily newspapers. *Econometrica*, *78*(1), 35–71.

Goldsby, T. J., Michael Knemeyer, A., Miller, J. W., & Wallenburg, C. M. (2013). Measurement and moderation: Finding the boundary conditions in logistics and supply chain research. *Journal of Business Logistics*, *34*(2), 109–116.

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, *78*(6), 1360–1380.

Hamaker, E. L., Asparouhov, T., Brose, A., Schmiedek, F., & Muthén, B. (2018). At the frontiers of modeling intensive longitudinal data: Dynamic structural equation models for the affective measurements from the COGITO study. *Multivariate Behavioral Research*, *53*(6), 820–841.

Hedström, P. & Ylikoski, P. (2010). Causal mechanisms in the social sciences. *Annual Review of Sociology*, *36*, 49–67.

Heide, J. B. & John, G. (1988). The role of dependence balancing in safeguarding transaction-specific assets in conventional channels. *Journal of Marketing*, *52*(1), 20–35.

Horowitz, K. J. & Planting, M. A. (2009). Concepts and methods of the US input–output accounts. Available at: www.bea.gov/sites/default/files/methodologies/IOmanual_092906.pdf.

Ibanez, M. R. & Toffel, M. W. (2020). How scheduling can bias quality assessment: Evidence from food-safety inspections. *Management Science*, *66*(6), 2396–2416.

Jin, G. Z. & Leslie, P. (2003). The effect of information on product quality: Evidence from restaurant hygiene grade cards. *Quarterly Journal of Economics*, *118*(2), 409–451.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527–535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319–342.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Kane, T. J. & Staiger, D. O. (2002). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives*, *16*(4), 91–114.

Ketchen, D. J., Ireland, R. D., & Baker, L. T. (2013). The use of archival proxies in strategic management studies: Castles made of sand? *Organizational Research Methods*, *16*(1), 32–42.

Lipton, P. (2004). *Inference to the Best Explanation*, 2nd ed. Routledge

Little, T. D. (2013). *Longitudinal Structural Equation Modeling*. Guilford Press.

Macher, J. T., Mayo, J. W., & Nickerson, J. A. (2011). Regulator heterogeneity and endogenous efforts to close the information asymmetry gap. *Journal of Law and Economics*, *54*(1), 25–54.

Mahoney, J. (2001). Beyond correlational analysis: Recent innovations in theory and method. *Sociological Forum* *16*(3), 575–593.

McKendall, M. A. & Wagner, J. A., III (1997). Motive, opportunity, choice, and corporate illegality. *Organization Science*, *8*(6), 624–647.

McKone, K. E. & Weiss, E. N. (1998). TPM: Planned and autonomous maintenance – bridging the gap between practice and research. *Production and Operations Management*, *7*(4), 335–351.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*(2), 108–141.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749.

Miller, J. & Parast, M. M. (2019). Learning by applying: The case of the Malcolm Baldrige National Quality Award. *IEEE Transactions on Engineering Management*, *66*(3), 337–353.

Miller, J. W. & Saldanha, J. P. (2016). A new look at the longitudinal relationship between motor carrier financial performance and safety. *Journal of Business Logistics*, *37*(3), 284–306.

Miller, J. & Saldanha, J. P. (2018). An exploratory investigation of new entrant motor carriers' longitudinal safety performance. *Transportation Journal*, *57*(2), 163–192.

Miller, J. W., Golicic, S. L., & Fugate, B. S. (2018). Reconciling alternative theories for the safety of owner–operators. *Journal of Business Logistics*, *39*(2), 101–122.

Miller, J. W., Muir, W. A., Bolumole, Y., & Griffis, S. E. (2020). The effect of truckload driver turnover on truckload freight pricing. *Journal of Business Logistics*, *41*(4), 294–309.

Miller, J. W., Bolumole, Y., & Muir, W. A. (2021a). Exploring longitudinal industry-level large truckload driver turnover. *Journal of Business Logistics*, *42*(4), 428–450. https://doi.org/10.1111/jbl.12235

Miller, J. W., Scott, A., & Williams, B. D. (2021b). Pricing dynamics in the truckload sector: The moderating role of the electronic logging device mandate. *Journal of Business Logistics*, *42*(4), 388–405. https://doi.org/10.1111/jbl.12256

Miller, J., Davis-Sramek, B., Fugate, B. S., Pagell, M., & Flynn, B. B. (2021c). Editorial commentary: Addressing confusion in the diffusion of archival data research. *Journal of Supply Chain Management*, *57*(3), 130–146. https://doi.org/10.1111/jscm.12236

Miller, J., Skowronski, K., & Saldanha, J. (2022) Asset ownership & incentives to undertake non-contractible actions: The case of trucking. *Journal of Supply Chain Management*, *58*, 65–91. https://doi.org/10.1111/jscm.12263

Muir, W. A., Miller, J. W., Griffis, S. E., Bolumole, Y. A., & Schwieterman, M. A. (2019). Strategic purity and efficiency in the motor carrier industry: A multiyear panel investigation. *Journal of Business Logistics*, *40*(3), 204–228.

Muthén, B. & Asparouhov, T. (2014). IRT studies of many groups: The alignment method. *Frontiers in Psychology*, *5*, 978.

Muthén, B. & Asparouhov, T. (2018). Recent methods for the study of measurement invariance with many groups: Alignment and random effects. *Sociological Methods & Research*, *47*(4), 637–664.

Nyrup, R. (2015). How explanatory reasoning justifies pursuit: A Peircean view of IBE. *Philosophy of Science*, *82*(5), 749–760.

Pawson, R. & Manzano-Santaella, A. (2012). A realist diagnostic workshop. *Evaluation*, *18* (2), 176–191.

Peltzman, S. (2000). Prices rise faster than they fall. *Journal of Political Economy*, *108*(3), 466–502.

Sauder, M. & Espeland, W. N. (2009). The discipline of rankings: Tight coupling and organizational change. *American Sociological Review*, *74*(1), 63–82.

Schwieterman, M. A., Miller, J., Knemeyer, A. M., & Croxton, K. L. (2020). Do supply chain exemplars have more or less dependent suppliers? *Journal of Business Logistics*, *41*(2), 149–173.

Scott, A. (2015). The value of information sharing for truckload shippers. *Transportation Research Part E: Logistics and Transportation Review*, *81*, 203–214.

Scott, A. (2018). Carrier bidding behavior in truckload spot auctions. *Journal of Business Logistics*, *39*(4), 267–281.

Scott, A. (2019). Concurrent business and buyer–supplier behavior in B2B auctions: Evidence from truckload transportation. *Production and Operations Management*, *28*(10), 2609–2628.

Scott, A. & Nyaga, G. N. (2019). The effect of firm size, asset ownership, and market prices on regulatory violations. *Journal of Operations Management*, *65*(7), 685–709.

Scott, A., Balthrop, A., & Miller, J. W. (2021). Unintended responses to IT-enabled monitoring: The case of the electronic logging device mandate. *Journal of Operations Management*, *67*(2), 152–181.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin.

Singer, J. D. and Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford University Press.

Skowronski, K. & Benton Jr, W. C. (2018). The influence of intellectual property rights on poaching in manufacturing outsourcing. *Production and Operations Management*, *27*(3), 531–552.

Steel, D. (2004). Social mechanisms and causal inference. *Philosophy of the Social Sciences*, *34*(1), 55–78.

Sutton, R. I. & Staw, B. M. (1995). What theory is not. *Administrative Science Quarterly*, *40*(3), 371–384.

Syverson, C. (2004). Market structure and productivity: A concrete example. *Journal of Political Economy*, *112*(6), 1181–1222.

Vegter, A., Taylor, J. K., & Haider-Markel, D. P. (2020). Old and new data sources and methods for interest group research. *Interest Groups & Advocacy*, *9*(3), 436–450.

Williamson, O. E. (2005). The economics of governance. *American Economic Review*, *95*(2), 1–18.

Winter, S. G., Szulanski, G., Ringov, D., & Jensen, R. J. (2012). Reproducing knowledge: Inaccurate replication and failure in franchise organizations. *Organization Science*, *23*(3), 672–685.

Wirth, R. J. & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*(1), 58–79.

Zhang, G., Browne, M. W., Ong, A. D., & Chow, S. M. (2014). Analytic standard errors for exploratory process factor analysis. *Psychometrika*, *79*(3), 444–469.

20 Qualitative Research Design[†]

Sinikka Elliott, Kayonne Christy, and Siqi Xiao

Abstract

The social world is fascinating – full of complexities, tensions, and contradictions. Social scientists have long been interested in better understanding the social world around us. Unlike quantitative research, that focuses on collecting and analyzing numerical data to make statistical inferences about the social world, qualitative research contributes to empirical and theoretical understandings of society by examining and explaining how and why people think and act as they do through the use of non-numerical data. In other words, qualitative research uncovers social processes and mechanisms undergirding human behavior. In this chapter, we will discuss how to design a qualitative research project using two of the most common qualitative research methods: in-depth interviewing and ethnographic observations (also known as ethnography or participant observation). We will begin the chapter by discussing the *what*, *how*, and *why* of interviewing and ethnography. We will then discuss the importance of interrogating one's underlying ontological and epistemological assumptions regarding research (and the research process) and the steps to follow in designing a qualitative study. We conclude the chapter by reviewing the different elements to consider when developing a qualitative research project.

Keywords: Qualitative Research, Interviews, Ethnography

Introduction: Ethnography and Interviewing

Qualitative research in the social and behavioral sciences examines the world by investigating how and why people think and act as they do. Qualitative research is indispensable because of its capacity to interpret meanings and generate or advance theories (Strauss, 1987). In this chapter we discuss two of the most used qualitative methods: in-depth interviewing and ethnography.

In-depth Interviewing

In-depth interviewing is a method of qualitative data collection that involves researcher(s) asking a series of direct, open-ended questions to interview participants about a certain topic. By asking people questions about their experiences and

[†] We would like to dedicate this chapter to the memory of Dr. Sinikka Elliott, who made a significant impact on our lives and the discipline of sociology at large. She modeled a feminist ethic of care in her teaching, scholarship, and mentorship, and we hope this chapter gives readers a glimpse of her wisdom and activism.

feelings, the researcher gains greater insight into the motivations, justifications, meanings, and other thought processes behind individual behaviors. In-depth interviews can be structured, semi-structured, or unstructured. During structured interviews, the researcher will develop an interview protocol, consisting of a series of questions, before entering the field, and will closely follow the interview protocol when interviewing research participants. Some researchers prefer structured interviewing because it ensures that there will be no differences in the types of questions asked to all research participants. While the structured interview often results in better interviews for employment, by asking few to no follow-up questions based on participants' responses, the researcher may miss out on valuable information.

Structured interviews are rare in qualitative research (Esterberg, 2002) since semi-structured interviews offer more flexibility by posing a series of broad questions (ideally no more than 7–10) with potential probes to be followed up on as well as asking impromptu questions during the interview. Semi-structured interviews offer a blend of structure and the ability to carefully probe and understand participants' experiences and worldviews. Unstructured interviews typically involve the researcher asking the participant one broad, open-ended question followed by a series of probes and follow-up questions. The rationale to this approach is that it invites and creates an opportunity for the participant to take the lead in directing the course of the interview that may improve the flow and depth of information and enable the interview to go in new and potentially unexpected directions.

Ethnographic Observations

Ethnographic (or participant) observation is a method of data collection in which researchers enter a setting or settings ("the field site") for the purpose of writing detailed field notes about what they observe. On-site observation helps researchers capture in-situ actions and interactions. There are different ways to design and conduct an ethnographic study. Participant observation can be covert, overt, or somewhere in between. Covert participant observation means that the researcher's identity is concealed, and the communities being studied do not know they are being observed. In overt participant observation, the communities being studied have full knowledge of the researcher's identity and objectives and have consented to participate in the study. Sometimes, researchers might purposefully disclose selective information while hiding other information about themselves or their research to avoid social-desirability bias, increase participants' willingness to directly discuss certain issues, or protect researchers' safety.

During participant observation, researchers develop techniques to "record" what they observe, including what they see, hear, smell, taste, do, and feel. Some researchers rely exclusively on memory, and some write brief jottings in the field to aid their memory. Deciding what approach is best for your research depends on a variety of factors including convivence, appropriateness of the environment, and personal preference. After each observation, ethnographers write extensive field notes capturing all relevant details and descriptions. Observations and field notes generally become more focused over time as the researcher begins to develop an analysis or interpretation of what they are observing (Emerson et al., 2011).

While ethnographic observations and in-depth interviewing are two different research methods, qualitative research projects often involve both, although one method may be more extensively pursued than the other. For example, a researcher who intends to primarily draw on ethnographic observations may decide to include both formal and informal interviewing to better capture the meanings people give to the setting and their actions in it. Similarly, researchers conducting interview studies will typically write field notes after each interview, describing the location of the interview, the interview participant, and other salient aspects of the interview that an audio recording alone cannot fully capture. These field notes offer crucial insights into the interview dynamics, help to contextualize the interview in space and time, and provide details of the participants' appearance and mannerisms that can help bring them to life. In this way, in-depth interviewing and ethnographic observations are not mutually exclusive but can be, and often are, used in tandem.

When to Use Interviewing and/or Ethnography

Whether you will use ethnography and/or interviewing in your qualitative study depends on your research question since different qualitative research methods will generate different insights into the topic of study. For example, in-depth interviewing is a generative method of data collection if you are interested in studying the meanings people give to things, their worldviews, how they talk about and make sense of their experiences, and/or if you're interested in uncovering the story behind something you observe during ethnographic research. The interactive nature of semi-structured and unstructured interviews allows researchers to talk through a research topic with their participants, ask them follow-up questions, and develop new questions based on their responses. On the other hand, ethnographic research is a generative method of data collection if you are interested in studying how people behave and interact with others in specific settings. Other considerations may also shape your decision regarding which method of inquiry to employ. Some of these considerations are practical, such as funding, time constraints, access to participants or settings, and unexpected circumstances (e.g., a global pandemic that prohibited much in-person data collection). Other factors include a researcher's passion for the subject matter and ontological and epistemological stances.

Reflecting on our own experiences as students, researchers, and instructors, we recognize a widespread discourse of treating qualitative research methods as toolkits in teaching, researching, and publishing. While we agree that the methods of qualitative research are excellent tools to capture individuals' experiences and examine social processes, the focus on methods as mere tools is misleading; it obscures the fundamental ontological and epistemological enterprise behind research design. Our goal in this chapter is to introduce qualitative research design using a coherent approach wherein the "toolkits" (i.e., the methods) and researchers' ontological and epistemological standpoints are united. In the next two sections, we will briefly introduce two dominant ontological and epistemological positions.

Ontology, Epistemology, and Methodology

Take a moment to reflect on the following questions: (1) What is your position on what can be known about the social world? (2) How do you believe we come to know what we know? The answers to these two questions shed light on your underlying ontological and epistemological assumptions (see Figure 20.1). Whether or not these assumptions are explicitly stated, all researchers subscribe to an ontology and epistemology, which in turn inform their research process (i.e., their methodology). The relationship between ontology, epistemology, and methodology is critical and should be considered before engaging in any type of research study – regardless of the method of inquiry (e.g., qualitative or quantitative) or research method (e.g., focus groups or content analysis). We will begin our discussion by closely examining what each of these terms mean in relation to the research design and process.

Ontology is a branch of philosophy that is primarily concerned with the nature of being and reality. Ontological issues are related to what is possible for humans to know about the social world. For instance, what is real? Who decides the legitimacy of what is real? Does reality exist independently from human perceptions and interpretations? How do researchers reconcile conflicting perceptions about reality? These beliefs about the nature of reality constitute a researcher's ontological assumptions. A researcher's ontology contains important ways of viewing the world that set the stage for ideas about what can be studied and the types of "truth" claims that can be made based on the findings from their research.

Epistemology is the study of knowledge – a theory of knowing. While ontological issues are concerned with the nature of reality, epistemological issues are interested in questions pertaining to how we come to know what we know. For example, what does it mean to "know" something? Where does "knowledge" come from? What is the relationship between the researcher and the researched (Varpio et al., 2017)? Should it be close and empathetic or distant and neutral, for example? What and whose voices are included or excluded in knowledge claims? Is our understanding of certain phenomena shaped by our background and identities (e.g., gender, race, ethnicity, sexuality, and age)? How do we judge and assess what kind of knowledge is valid and reliable?

A range of ontological and epistemological positions exist. For the purposes of this chapter, we will take a closer look at two dominant philosophical positions: positivism and constructivism. Positivism is concerned with uncovering objective

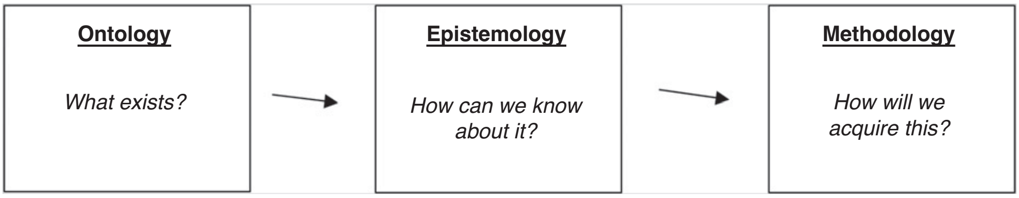


Figure 20.1 *Relationship between ontology, epistemology, and methodology.*

truths about the social world. A positivist ontological position subscribes to the belief that there is one single reality "out there," which we can gain access to through impartial, unbiased, and value-free scientific research methods. A positivist epistemological position believes that researchers should be completely objective to discover absolute truths about the social world.

Constructivism reflects very different underlying ontological and epistemological assumptions. Constructivists believe that individuals' personal perceptions and interpretations shape the truths we construct about the social world. A constructivist ontological position subscribes to the belief that multiple realities exist because reality is socially constructed by humans in different social contexts and under differing social conditions. Since reality is dependent on the interaction between humans and the social world, reality is subjective, differently interpreted, and constantly negotiated. In this sense, researchers are part of the social world and inextricably part of the research they do and the data they collect. A constructivist epistemological position believes that researchers cannot remove themselves from the research process – the researcher's social position, beliefs, and values influence all aspects of the research process from their interactions in the field to how they analyze and write up their findings.

To further elaborate on the critical role of ontology and epistemology in research, consider the concept of epistemic injustice, which alerts us to the ways dominant assumptions often infuse these traditions. Epistemic injustice, a term coined by Miranda Fricker (2007), calls attention to how individuals "can be unfairly discriminated against in our capacity as a knower based on prejudices about the speaker, such as gender, social background, ethnicity, race, sexuality, tone of voice, accent, and so on" (Byskov, 2020, p. 1). While epistemology is deeply concerned with how the knower sees and makes sense of the world, not all social groups are equally regarded as knowers in the academy (Collins, 2000; Todd, 2016). For instance, take a moment to reflect on your educational experiences as a student. Whose ways of knowing are more privileged in your lectures and assigned readings? Who commonly holds the authority to make knowledge claims and where does this authority come from? In social and behavioral science, disciplinary canons overwhelmingly consist of scholarship from "the founding fathers" – long deceased white European men (Morris, 2015; Sprague, 1997). These early theorists have contributed much to their disciplines; however, when academics prioritize the epistemologies of privileged groups over other social groups, this perpetuates epistemic injustice. In recent decades, those who have been historically excluded from the academy have made important epistemic contributions to advance knowledge outside of disciplinary canons (Collins, 1989; Tuck et al., 2014).

The Black feminist sociologist Patricia Hill Collins, for example, developed a body of knowledge around Black feminist thought drawing on Black women intellectuals, including blues singers, novelists, poets, activists, and working-class women. Collins (1989; 2000) eschewed the sociological canon because of the way it commonly pathologized Black women as well as the way it studies oppressed groups – as less than human and less capable of developing independent interpretation or articulating their standpoints. According to Collins (1989, pp. 747–748), "Black women's political and economic status provides them with a distinctive set of experiences that offers a different view of material reality than that available to other groups . . . these experiences stimulate a distinctive Black feminist consciousness concerning that material reality." Four dimensions of an Afrocentric feminist epistemology underpin Black feminist thought: (1) concrete experience as a criterion of meaning; (2) the use of dialogue in assessing knowledge claims; (3) the ethics of caring; and (4) the ethic of personal accountability (Collins, 1989; 2000). In this way, Black feminist thought not only offers distinctive ways of knowing about the social world and challenges the dominant Eurocentric masculinist epistemologies but also has profound implications for how we produce and evaluate knowledge. As Black feminist thought and epistemic injustice reminds us, scholarly mechanisms of knowledge production and validation stem from and reflect various ontological and epistemological positions – some of which are rooted in historical and ongoing inequalities and injustices that, when left unexamined, are reproduced.

As discussed earlier, a researcher's ontological and epistemological assumptions shape their methodology. Methodology is a system of broad principles or rules that underpin the specific methods or procedures a researcher uses to reveal and explain the phenomena of interest. Methodology informs many decisions researchers make during a research project – *how* and *why* we pose questions, collect evidence, and analyze data – because it forms the guiding principles behind the research. If you subscribe to the notion that valid research involves the researcher being value-free, and merely reporting the data (i.e., positivist), as an ethnographer you may believe that it is best to only observe and refrain from participating in your field site in order to remain objective and avoid "contaminating" the data. You may believe that becoming too involved in the field will jeopardize your findings and, thus, the generalizability and replicability of the study. In contrast, if you value research that positions knowledge as socially constructed, a product of the relationship between the researcher and that being researched (i.e., constructivist), as an ethnographer you will seek to participate actively in your field site while constantly reflecting on what your presence means for what you are observing and how your experiences in the field provide important insights. To reiterate, underlying philosophical assumptions are critical to consider while conducting research.

We have barely scratched the surface of these important yet esoteric-seeming issues. We recognize that this discussion may seem abstract to some readers. We raise these ontological, epistemological, and methodological issues to remind researchers that methods are not simply the procedures researchers follow to gather and analyze data. Behind the methods we utilize are many assumptions about what is knowable, what is worth knowing, how we can know it, and how we should study and report it. Well-designed qualitative research projects need to have a coherent ontology, epistemology, and methodology, and it is essential that researchers examine their own stance on these assumptions prior to, and while, undertaking a qualitative project. To be transparent, we locate our work in the constructivist tradition, which perceives knowledge as co-created by and through social relations, contexts, and power dynamics.

Developing a Research Question

Qualitative research projects often start when a researcher observes something puzzling or curious in the social world or “the literature.” The process of reviewing and synthesizing the existing literature on a research topic, namely, a “literature review,” is discussed in more detail in Chapter 4 of this volume. Literature reviews often inspire new qualitative research projects. Unexplained puzzles or contradictory findings in the literature are often at the heart of excellently designed qualitative research studies. You may also find a methodological gap in the literature – most studies have taken X approach but there is reason to think Y approach would yield important insights into a social phenomenon. Qualitative researchers, however, should be cautious about proposing a simple “gap in the literature” as a reason for doing a qualitative project. Just because no one has studied a particular group or issue is not a compelling rationale for a qualitative project. Instead, you will need to make a case for why existing theories of human behavior and the social world will be augmented by your proposed study.

Most qualitative research adds to or extends the extant literature by examining “a previously ignored sub-population, a different time frame, or an event that may have affected the group or organization of interest” (Aurini et al., 2016, p. 28). However, Janice D. Aurini and colleagues caution that simply “adding a new case does not automatically make for an interesting research problem . . . *You must first articulate why the new case is a meaningful extension of the literature* . . . ” (Aurini et al., 2016, p. 28–29, emphasis in original). In other words, be prepared to answer the “so what” question. Why should anyone care about this research project? What is it about your project that is going to contribute new and necessary knowledge?

Developing a research question is crucial in the beginning stages of any research. The question or questions that you pose will guide every element of your research design. As you mull over a potential topic for a qualitative study, ask yourself what the issue or puzzle to be addressed through this research is. What data do I need to collect to provide answers to this problem or puzzle? What is the best method(s) to use to collect this data? Qualitative research involves collecting empirical data, and thus, qualitative research questions should be grounded in “the empirical world” (Esterberg, 2002, p. 30). Your research question should also define the parameters of the study. For example, say you want to study racism. What is it about racism you wish you examine? Do you want to know about the lived experiences of a particular racialized group within the healthcare system? Do you want to examine how racialized mothers made sense of food programs for their children at school? Perhaps you want to know how the news media discussed racial disparities in online dating? Each of these studies would offer insight into racism, albeit in different ways. If you find yourself writing abstract research questions that do not ground your focus in the empirical world, this could be an indication that you have not yet figured out the angle or focus you wish to take. Drafting up several questions about the phenomenon you wish to study, which are concretely anchored in elements of the empirical world, can help you identify the specific research question, and hence project, you wish to pursue. Your research question should point to the data you will need to collect and analyze to answer the question(s) you've posed.

In addition to being attuned to the empirical findings or gaps in previous research, the review of the literature for qualitative research should be conceptual – how have others conceptualized the problem? What concepts (or theories) have been developed to explain the phenomenon? What conceptual gaps exist in the literature? A novice researcher may immediately feel overwhelmed by the volume of research on their topic of interest. One way to handle this information overload is to first seek out review articles that overview the state of the literature on your topic of interest and propose future directions it should take. Also look for theoretical pieces that examine how the issue has been theorized and lay out ideas for new theoretical advances. You should pay attention to key words that have been regularly cited and reflect on what we learn from this focus and what might be missing from it. Remember, in doing a literature review for a qualitative project, you are trying to identify a gap in knowledge about the mechanisms and processes underlying a particular phenomenon, which your proposed research would be able to fill (Small, 2009).

One way to conceptually ground a qualitative project in the literature is to begin with what Herbert Blumer (1969) termed "sensitizing concepts." Examples of sensitizing concepts include "feeling rules" (Hochschild, 1979), "intensive mothering" (Hays, 1996), or "intersectionality" (Crenshaw, 1991). This approach is advocated by Kathy Charmaz (2014), who advances a constructivist grounded theory approach to qualitative research. Rather than avoiding the literature at the onset of a qualitative project, as Glaser and Strauss (1967) advocated in their classic treatise on grounded theory, Charmaz advises researchers to use "those [sensitizing] concepts as *points of departure* to form interview questions, to look at data, to listen to interviewees, and to think analytically about the data" (Charmaz, 2014, p. 31, emphasis in original). Ultimately, your goal is to contribute to scholarship on your topic. You may have other goals as well, but in academia making contributions to the literature is considered de rigueur. It makes sense, therefore, to know what others have to say about the phenomenon before conducting your project.

Qualitative research projects require a blend of both careful design *and* flexibility. Thoughtfully constructed research questions often shift once data collection begins. Jessica McCrory Calarco (2018) initially set out to study cross-class friendships in an elementary school consisting of middle-class and working-class students. Over the course of observing school interactions during a two-year ethnography, she noticed a pattern in her data – middle-class students took various actions, often involving asking teachers for help, "to overcome problems that stymied their working-class peers" (Calarco, 2018, p. 2). Her focus thus shifted "to ask: *How does the middle-class secure unequal advantages in school?*" (Calarco, 2018, p. 2, emphasis in original). As was the case for Calarco, you should anticipate that you will need to modify your research question over the course of your qualitative study (Luker, 2008). Nevertheless, it is essential to start with a research question that grounds and focuses the initial stages of data collection.

Sampling

You want to observe and learn from individuals and settings where you expect to find the phenomenon of interest. Yet, given the way the social world works, there is good reason to expect that the phenomenon you wish to study will vary in different settings and in different contexts. Sampling in interview studies and ethnographies refers to the process of selecting a unit of analysis to investigate (e.g., an organization, a family, and individuals). Qualitative researchers must think carefully about whom they will interview and/or what they will observe to answer their research question(s). The research question is critical for initially guiding the selection of the interview sample or field site(s). For example, your research question may ask whether and how the phenomenon of interest varies for different people or in different settings. This focus suggests a comparative study in which you "zero in on the groups that will foster strategic comparisons" (Gerson & Damaske, 2021, p. 27). Going back to our previous example of studying racialized groups' experience of romantic relationships, if we want to study East Asian women's experience specifically, we can suspect that individuals' experience might differ depending on their gender, sexuality, age, immigrant background, etc. If you suspect that the phenomenon of interest may vary for different subcategories of a group, then you will need to develop eligibility criteria to screen and assess (or "filter") a potential field site or interview participant, to determine whether they fit into one of the subcategories you have established. You might decide, for instance, that participants for your interview study must fit into certain demographic parameters that you have established.

Another approach to sampling is to strategically build a diverse sample to "claim that the sample included the full variety of instances that would be encountered anywhere" (Weiss, 1994, p. 24). In sampling for range (Small, 2009; Weiss, 1994), researchers "select respondents purposively so that we obtain instances of all the important dissimilar forms present in the larger population" (Weiss, 1994, p. 20). Sampling for range does not mean that the researcher can claim that the sample is representative of the general population; it is intended to help the researcher build a robust theory of human behavior by including a wide variety of experiences and/or interactions rather than hearing or observing the same thing over and over again.

In contrast to sampling for range, your sampling strategy might be based on locating an extreme group or situation (i.e., an extreme case), where you expect to find the phenomenon of interest is heightened (Williams, 1991). Again, you would need to establish the parameters of the sample – the criteria you will use to establish that something counts as an extreme case. It might be a group where you expect the phenomenon is fervently held, policed, or repudiated. In selecting the sample, you are looking for cases (e.g., individual identities and organizational settings) that involve deep investment, contradictions, ambivalences, double binds, and so on. Along these same lines, another sampling strategy is to find a unique identity, group, or setting to examine (i.e., a unique case; Small, 2009). Unique cases defy stereotypes or general patterns found in the social world. In following this sampling strategy, your task is to understand and explain the unique case's deviation from the norm. By explaining why the case does not conform to the norm, the answer will shed important light on the unique case and on the norm (Small, 2009).

Snowball sampling is a widely used practice for recruiting interview participants. In building a snowball sample, the researcher asks participants to refer them to other participants. Snowball sampling has the advantage of "increas[ing] the number of respondents, because people become more receptive to a researcher when the latter has been vouched for by a friend as trustworthy" (Small 2009, p. 14). Although snowballing can result in a sample of people who may form a social network, this should not be seen as a flaw or bias of the sample; rather, it is a particular characteristic of the sample that "should be understood, developed, and incorporated into [the researcher's] understanding of the cases at hand" (Small, 2009, p. 14). Researchers employing snowballing, who hope to construct a diverse sample, can ask participants to refer them to individuals who are dissimilar from them, such as those who may have very different experiences with or thoughts on the phenomenon of interest.

Theoretical sampling is another form of sampling in qualitative research (Charmaz, 2014). In theoretical sampling, qualitative researchers develop a theory of what they are learning in the field, during data collection, and then collect more data to elaborate and refine the nascent theory. This might involve going back into the same field setting(s), re-interviewing the same people, conducting more interviews with new people, or adding a new field site to the study to further develop and refine the theory. As the practice of theoretical sampling implies, researchers cannot always anticipate, in advance, what they are going to find once they start collecting data. Thus, in addition to beginning your project with a carefully constructed design, you should be attuned to what you are learning in the research and be prepared to modify your study, if necessary (e.g., by adding an additional case, revising the research question(s), or modifying your sampling strategy).

How many interviews or how many observations are necessary to do is often on the minds of researchers as they design qualitative studies (Small, 2009). Depending on your ontological and epistemological commitments, you may develop quotas to ensure that you interview a certain number of people in each of your subcategories or observe a certain number of events in the field. Alternatively, you may decide how many interviews or observations you will conduct based on the principle of analytic saturation – you stop collecting data once you have fleshed out your theory in full (Small, 2009).

Because qualitative studies are not intended to be representative of a population, qualitative researchers defend their sample based on how well it allows them to uncover mechanisms and trace processes (Small, 2009). Did we get full access to all aspects of the field site to develop a robust explanation of what was happening? Did we build trust and rapport such that interview participants talked in great depth and detail about the issues they face and how they make sense of them? The goal of qualitative research is to answer core questions underlying social behavior, such as "[B]y what *process* do outcomes develop, and what *factors* and *mechanisms* influence their emergence?" (Gerson & Damaske, 2021, p. 164, emphasis in original). Answering these core questions requires gathering rich, contextually embedded data; gaining access to the people and places that will engender this type of data is crucial.

Gaining Access

Once you've established your research question and have decided what data you need to collect to answer it, you must decide how you are going to get access to those data. For an interview study, this involves figuring out how you are going to find and tap into individuals who meet the eligibility criteria you have established and will agree to be interviewed. For an ethnographic study, you will need to figure out the potential field site or field sites (for a multi-sited study) that meet the parameters you have set and how to gain access to them. Many qualitative studies involve both interviews and ethnographic observations; in this case, the researcher will need to both gain access to a field site and recruit people for interviews. Some qualitative projects that involve collaboration with community partners, often referred as community-based participatory research, may also involve gaining access to organizations and a range of community stakeholders (Banks et al., 2013).

Gaining access to interview participants and field sites is exciting and nerve wracking. Seasoned qualitative researchers often stress how crucial this aspect of a qualitative study is both because your study cannot proceed if you do not get access and because *how* you get access shapes the data you collect (Luker, 2008). For example, how you describe your study to potential participants and in social media posts, flyers, and other recruitment materials, and where you post them, can influence who responds to your call for participants in an interview study. Additionally, how you build trust and rapport with your informants when gaining access to the field sites or community partners can influence how participants share their stories.

It is important to think carefully about how you are going to frame your study to your potential research population and come up with a convincing hook (Luker, 2008). Why should they give you access to their group or organization? What would motivate someone to take time out of their day to be interviewed or agree to allow you into their group or organization to be observed? As Kristin Luker puts it, "your task is to figure out what's in it for others to participate" (Luker, 2008, p. 147). You must convince them that there is something important and compelling that you wish to study and that they have great insight into it. Try to pique their interest. Luker advises framing your research project in ways that resonate with those you hope will participate. One way to do this is to find out what is on the minds of people who fit your study criteria. This might involve conducting pilot interviews that both test out your interview guide and help you tap into what matters to the individuals you intend to study – thus helping you to strategically frame your research project to them.

To help answer his research question on how states govern urban poverty in the United States in the twenty-first century, Armando Lara-Millán (2021) had to figure out how to get into organizations to observe the operations of the state in action. By interviewing high-profile members of organizations and then using their way of talking about the problems they faced to frame his research interests, Lara-Millán was able to gain access to a large urban jail and a public hospital where he conducted his ethnographic study. For community-based research, particularly, you might discuss with potential community partners about how the communities would benefit from a collaboration (e.g., a reciprocal exchange of expertise or co-learning opportunities; Banks et al., 2013; Coughlin et al., 2017).

Qualitative researchers sometimes get access to a field site or interview population through a main informant; that is, someone on the "inside" who vouches for you, helps you get in, and introduces you to those you wish to study. In this way, a main informant both helps you gain access and establish trust and rapport in the field. This method of access is most common in ethnographic studies, but it can also be a starting point for an interview study. Snowball samples, discussed earlier in the chapter, often start with a main informant who connects the researcher to two or three interview participants, who then refer other potential interview participants.

Access is not a one-time occasion, in a qualitative project, but rather has to be continually negotiated. You may gain access to a field site, but those in it may treat you as a distant outsider until they have decided if you are trustworthy, capable of understanding their lives, and will not misrepresent them (Jones, 2010). Interview participants may be reluctant to open up or even cut short the interview if they do not trust your motives or if your affect or line of questioning makes them uncomfortable. Qualitative research requires asking questions and observing goings on that, without gaining trust and rapport, might seem intrusive and put people on guard. You should anticipate managing a delicate balance between building rapport and trust while digging deep to try to gain access to complex, behind-the-scenes actions, interactions, and meanings.

Designing an Interview Guide

The broad goal of conducting an interview study is to learn why people do what they do from their point of view. Qualitative interview questions should be open-ended and invite participants to talk about their experiences, feelings, and views, ideally without judgment. Avoid posing questions that elicit yes or no responses as these can set up a more survey-like, rather than in-depth interview, atmosphere. For instance, instead of asking "Do you like your job?," an in-depth interviewer might ask "What do you think about your job?" *Qualitative interviews also try to avoid asking leading questions.* These are questions that may lead a participant to think about an issue in a certain way. For example, rather than asking "Did you eat the doughnuts because you were feeling stressed?," you might ask, "What were you feeling when you ate the doughnuts?"

The qualitative interview uses two kinds of questions: main questions and follow-up questions. Main questions begin and guide the interview. They should be organized around the main theoretical or empirical aims of the study – they should help you answer the research question(s) you have established. An interview guide groups the main questions thematically so that they flow comfortably, while giving interviewers permission to organically stray from the guide during interviews. Sometimes you may need to include a transition to a new topic in the guide (e.g., "These next few questions focus on . . . "). Follow-up questions flesh out the details of answers to main questions, such as asking for clarification ("What did you mean by X?"), requesting examples ("Can you give me an example of a time when that happened?"), and getting other details, including feelings about incidents ("Where were you when you heard that?," "Who else was there?," and "How did that make you feel?").

Often, interviewers include specific "probes" as follow-ups to main questions in the interview guide. These probes are in the guide to remind you of the various dimensions of the main question that you want to be sure to cover. Whether a main or a follow-up question, questions should be posed to elicit concrete information and avoid narrowly framing the issue or making the participant feel as though they are being judged. For example, qualitative interviewers often avoid starting a question with the word "Why" since this can lead people to feel as though their judgment is being questioned or they are being assessed. Instead, start questions with phrases such as "Walk me through ... " or "I'd like you to tell me in as much detail as possible ... ." The purpose of qualitative interviews is to both collect people's experiences and their interpretations of those experiences – how they subjectively experience and make sense of specific aspects of their social worlds. To accomplish this, Robert S. Weiss (1994) recommends focusing on areas that generate concrete descriptions rather than overgeneralized accounts.

The questions you ask shape the answers you get. Researchers working in the positivist tradition will try to avoid biasing the interview through their questions to gain an "uncontaminated" understanding of their phenomenon under investigation. In the constructivist tradition, the researcher is active in shaping the data. As Weiss (1994, p. 65) puts it, "the interviewer and the respondent will work together to produce information useful to the research project." In this vein, interview participants are not simply "vessel[s] waiting to be tapped" with skillful, neutral questions; instead, their "interpretive capabilities must be activated, stimulated and cultivated" (Holstein & Gubrium, 2002, p. 120). The key task for researchers in this approach is to pay attention to how their questions helped to co-construct knowledge.

During qualitative interviews, the goal is to gain the participants' trust and build rapport so they open up to you. You will need to think carefully about how you plan to achieve this during the design stage of the project. Novice interviewers sometimes rely on praising or agreeing with the participant to build trust and rapport. However, many qualitative researchers argue that interviewers should avoid appearing to agree or disagree with participants; if interview participants think the interviewer is playing an evaluative role, they might not disclose information that would cast them in a less favorable light. Rather than responding evaluatively ("It's great that you won the award!"), interviewers often mirror back what a participant has said to acknowledge they have heard it and open space for the participant to further elaborate ("You won the award. What did that mean to you?"). A major aspect of in-depth interviewing is for participants to explain what things mean to them. Even if you know the meaning of something (or think you do), it is important to ask so that you have it "on tape."

After each interview, qualitative researchers write field notes that include information about the time and place of the interview and a description of the participant and anything they might have said before or after the recorded interview that provides relevant information (e.g., insight into how they were viewing you and what they thought of the interview). Interview field notes should also summarize what the participant said and highlight the main analytic themes from the interview. Finally, field notes may include feelings or reflections the interviewer had during or after the interview and suggest avenues to explore in future interviews, including possible changes to the interview guide.

Writing Ethnographic Field Notes

Ethnography is ideal for a study aimed at contextualizing people's actions, interactions, and subjectivities. Specific, vivid details are essential for capturing the social contexts that shape people's experiences, identities, motivations, and so on. Ethnographers rely on painstakingly written field notes to document what they observe and experience in the field – what Emerson and colleagues call "*descriptive* fieldnotes" (Emerson et al., 2011), p. 5, italics in original). In their comprehensive guide to writing field notes, these authors encourage ethnographers to avoid evaluative terms in their field notes and instead to carefully describe what they observe. Rather than saying a home is "messy," for example, flesh out the details that led you to form this impression (e.g., dirty laundry on the floor, a layer of dust on the furniture, and toys scattered about). The maxim "show don't tell" is relevant to field-note writing; ethnographers should show, in lush detail, in their field notes what they observed rather than trying to explain "*why* events or actions occur" (Emerson et al., 2011, p. 27, emphasis in original). "Focusing on *how routine actions in the setting are organized and take place*" (Emerson et al., 2011, p. 27, emphasis in original) rather than why allows you to document the processes by which people create, reproduce, and alter their social worlds.

Emerson et al. (2011, p. 6) stress that "there is no one 'natural' or 'correct' way to write about what one observes." Field-note descriptions inevitably "involve issues of perception and interpretation" because different fieldworkers have "distinctive orientations and positionings" (Emerson et al., 2011, p. 6). Thus, when designing an ethnographic study, you should reflect on the choices you are going to make in representing what you observe and how they align with the aims of the study. If you are interested in how spaces shape social interaction, for example, you will focus a great deal on the spatial aspects of what you observe. If you want to better understand the role of emotions in specific settings, then your focus will deeply attend to the emotional dynamics and tenor therein. The social world is infinitely interesting, and you may find that your focus changes once in your field site. Nevertheless, having an initial focus helps to hone your lens and reminds you of the analytic purpose of conducting fieldwork.

How you get access to field settings, as well as how and where you position yourself in the field, shapes what you see and are privy to. You should make decisions about when to be in the field and how to position yourself once in the field based on what social practices you're interested in and when and where they will be most likely to appear. Be prepared to develop strategies for connecting with a variety of individuals in your field site to seek out different experiences and viewpoints. Other methodological considerations include how to present yourself in the field, including how you describe your research interests, how you dress, and other aspects of your appearance.

You will also need to decide how you will record information in the field. Fieldworkers often write jottings in a small notebook (or phone) while in the field that they refer to as they write their field notes. Jottings entail “details of what you sense are key components of observed scenes, events, or interactions” (Emerson et al., 2011, p. 31), such as snippets of conversations, drawings to capture the layout and people’s placement within it, and “fragments of action” (Emerson et al., 2011, p. 31). Fieldworkers who do not write jottings while in the field sometimes furiously jot down keywords and other relevant pieces of information as soon as they leave the field to support the subsequent writing of field notes. Plan to write field notes as soon as possible after each time in the field and dedicate a significant amount of time to writing them. Generally, you can expect to spend two to three times as much time writing field notes as you spent in the field. Therefore, when designing your study, be sure to factor in sufficient time for writing extensive field notes.

Positionality and Reflexivity

Throughout the design of your project, you should examine your positionality – your own identity investments, experiences, and ideas in relation to those you wish to study and learn from. Researchers use “reflexivity” to refer to the process by which they reflect on their social position in relation to the field. Some questions to ask yourself include: Why do I want to do this study? Can I handle the physical, emotional, temporal, and other demands it will require of me? What are my experiences in relation to the phenomenon I wish to study? How do my experiences shape my ideas and interpretations? How might people in my study “read” me, and what are the implications of this for what they are willing to share with me? We recommend keeping a journal that starts by answering these questions and continues to reflect on your positionality throughout the study.

You will also need to consider other personal elements as you design your study, such as how much you intend to share about yourself during the research. For example, do you disclose details about your own life and worldview and share your own experiences with your research participants? Some qualitative feminist researchers argue that this is a way to gain greater insight and reduce some of the power relations in research (Carpenter, 2005; Oakley, 1981). However, others note that these practices pose their own ethical and practical dilemmas and question whether there can be a feminist ethnography (Stacey, 1988). On the whole, qualitative researchers need to be attentive to what we take into research situations (i.e., our backgrounds and expectations that form our embodied locations) and take out of them (i.e., the data we collect and our interpretations of them). Doing so necessitates being attentive to power relations, including our position in hierarchies of power and privilege (Sweet, 2020; Qin, 2016), and how this shapes the data we collect and the insights we form from them.

From the outset of a qualitative study, you should consider how you will position yourself within the research itself. Issues of positionality and reflexivity do not end once the researcher has completed data collection and analysis, however. Will you write yourself into and be part of the written reports (e.g., including your feelings and experiences) or will you avoid this and present the data as if you were not a part of them? As put by Qin (2016, p. 1), "researchers are always positioned but the disclosure of that positionality has not always found its way into the final research process." From a constructivist standpoint, the decisions researchers make in the field and the experiences they have form part of the data they collect and are subsequently included in written reports. For example, when C. J. Pascoe (2011) conducted fieldwork as a female researcher at River High, a suburban high school, male high-school students sometimes hit on her. Pascoe included these interactions as part of the data she was collecting and reflected on how they helped her to understand the processes she was observing: "As a female researcher I was drawn into a set of objectifying and sexualizing rituals through which boys constructed their identities and certain school spaces as masculine. In the end, I wasn't just studying their gender identities; I became part of the very process through which they constructed these identities" (Pascoe, 2011, p. 176).

Issues of positionality and reflexivity also concern how you write about and characterize your participants and field site(s). You will have captured a great deal of descriptive details in your research that, in subsequent published writing, help to bring readers into the scene and to connect with your participants (Emerson et al., 2011). However, you should carefully consider whether details are relevant to your argument and the story you are telling prior to sharing them with an audience. Sometimes researchers include details in published reports that are tangential and may be added to burnish their reputation (e.g., by demonstrating how close to the action they got) or to titillate the reader (Small, 2015).

You may face institutional pressure to include salacious details and should think through a rationale for why you will or will not do so. For example, when asked by editors to include more information about the sex lives of the poor Brown and Black youth she studied, Ranita Ray (2017) refused. Her rationale was that this information was not relevant to the arguments of the book and including it would be epistemically unjust given the ways elites have used depictions of people of color as hypersexual to justify their exploitation and oppression (Ray, 2021). Nikki Jones (2010) discusses in the methodological appendix of her book, *Between Good and Ghetto*, that readers often commented on how matter-of-factly she presented the lives of the inner-city Black girls in the book, revealing that what is expected from urban ethnographies is sensational details about the lives of inner-city residents (Small, 2015). Because qualitative researchers collect, analyze, and write up the research, they are inextricably a part of it. Designing a qualitative study requires developing explicit processes that enable the researcher to reflect on how they are situated in the data and research process.

Doing Ethical Research

The broad ethical precepts of research are to do no harm to those who participate in your research and are represented by it, your profession, institutional affiliation, and yourself (see Chapter 2 in this volume). To conduct a qualitative study, you will need to get ethical approval for the project from your institution's ethics board (or institutional review board). For qualitative research that involves interviews and ethnographic observations, this typically entails an application that includes your research proposal, recruitment materials, consent forms, interview guides or other data-collection protocols, and a detailed description of the steps you will take to ensure the confidentiality of the data you collect (e.g., encrypting data, de-identifying interview transcripts and field notes, and assigning pseudonyms to participants). There are concerns pertinent to the assumptions regarding the power dynamics between the researcher and participants in institutional ethical review processes, which do not fit with community-based research (e.g., what are the roles of participants and who owns the data; Banks, et al., 2013; Manzo & Brightbill, 2007). If you are collaborating with community partners, you might consider establishing a community advisory board consisting of community members and representatives from organizations, to help keep the research process ethically accountable as the partnership evolves, and revisit partnership agreements constantly (Banks et al., 2013). If your study poses risks to participants, you will need to justify why the research should still be carried out.

Simply having your research approved by an ethics board and following the agreed upon procedures does not guarantee that you will conduct an ethical project. You will likely encounter many ethical issues while conducting your study that you did not anticipate and may not clearly know how to resolve ethically (Stacey, 1988). As suggested by Aurini and colleagues, "[g]enerally, issues arise because qualitative researchers work with participants face-to-face, over long periods of time, and possibly in intimate circumstances. There can be a fine line between building relationships that are caring and not exploiting participants" (Aurini et al., 2016, p. 59). You will ultimately have to decide what that fine line is and develop justifications for the line you draw.

One of the ethical obligations of researchers is to avoid deception. One way to do this is through informed consent, whereby you provide participants "with information about the study's purpose, funding, the research team, how data will be used, and what will be required of them" (Aurini et al., 2016, p. 59–60) and ask them to voluntarily consent to participate in the study (see Chapter 10 in this volume). When participants know what they are agreeing to, they will not be surprised by the type of questions you ask or the degree of access you hope to gain in the field site. In this way, informed consent helps qualitative researchers by preparing participants for what the study entails. Although you should avoid deceiving your participants, in a qualitative study you do not always know from the outset what you are going to learn from the research that might take the study in new and unanticipated directions. Thus, while you should be as faithful as possible in describing your study to participants, the final research question(s) and study may look very different from the one you set out to do.

Most qualitative research projects promise participants that they will never be named in any presentations or written texts to come out of the study – this is called maintaining confidentiality. Qualitative researchers also typically change and conceal details about participants or the study setting that might be identifying. One exception might be a situation in which participants are very well known, making it next to impossible to conceal their identity. In this case, participants might be asked if they are willing to have their identities disclosed or informed that, despite efforts to conceal their identities, others will likely know who they are. Sometimes, it is hard to maintain confidentiality because you are studying people who know one another and thus, despite your best efforts to de-identify the data, may be able to recognize one another in the published report. In this case, you should inform participants of this possibility before they agree to participate. Confidentiality is one of the key tenets of qualitative research. In designing a study, you will need to develop thoughtful procedures to obscure identifying information and protect the information you have collected from data breaches.

Conclusion

Qualitative research unlocks crucial insights into the mechanisms and social processes – the how and why – of human behavior; it involves both philosophical and practical issues. Existing qualitative research stems from a variety of ontological and epistemological stances (Esterberg, 2002; Small, 2009). As a qualitative researcher, you will need to determine what approach you intend to pursue at the outset of your project. This will help you maintain methodological consistency throughout the study. Two dominant approaches that we have discussed in this chapter are positivism and constructivism. Your underlying ontology and epistemology shape many decisions you will make during the study and how you will defend those decisions in making claims about the validity of your research.

Thoughtfully designed qualitative research entails both rigor and flexibility. Research questions are essential to anchoring a qualitative project but often change during a study as the researcher hones in on the key analytic issue or puzzle to be addressed. Interview guides and field-site protocols may shift as you gain a better understanding of what is happening and important in the lives of others. Gaining access to the people and sites you wish to learn from, reflecting on how to position yourself within the research, and working through the many ethical issues that stem from enmeshing yourself in and reporting on the lives of others are critical aspects of qualitative research design. In all, carefully and thoughtfully constructed qualitative research has much to contribute to knowledge about the social world.

References

Aurini, J.D., Heath, M., & Howells, S. (2016). *The How to of Qualitative Research*. SAGE Publications.

Banks, S., Armstrong, A., Carter, K., et al. (2013). Everyday ethics in community-based participatory research. *Contemporary Social Science*, *8*(3), 263–277.

Blumer, H. (1969). *Symbolic Interactionism: Perspective and Method*. Prentice Hall.

Byskov, M. F. (2020). What makes epistemic injustice an ‘injustice’? *Journal of Social Philosophy*, *52*(1), 114–131.

Calarco, J. M. (2018). *Negotiating Opportunities: How the Middle Class Secures Advantages in School*. Oxford University Press.

Carpenter, L. (2005). *Virginity Lost: An Intimate Portrait of First Sexual Experiences*. New York University Press.

Charmaz, K. (2014). *Constructing Grounded Theory*, 2nd ed. SAGE Publications.

Collins, P. H. (1989). The social construction of Black feminist thought. *Signs*, *14*, 745–773.

Collins, P. H. (2000). *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*, 2nd ed. Routledge.

Coughlin, S. S., Smith, S. A., & Fernandez, M. E. (2017). Overview of community-based participatory research. In S. S. Coughlin, S. A. Smith, & M. E.Fernandez (eds.), *Handbook of Community-Based Participatory Research* (pp. 1–10). Oxford University Press.

Crenshaw, K. (1991). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, *43*(6), 1241–1299.

Emerson, R. M., Fretz, R. I., & Shaw, L. L. (2011). *Writing Ethnographic Fieldnotes*, 2nd ed. University of Chicago Press.

Esterberg, K. G. (2002). *Qualitative Methods in Social Research*. McGraw Hill.

Fricker, M. (2007). *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press.

Gerson, K. & Damaske, S. (2021). *The Science and Art of Interviewing*. Oxford University Press.

Glaser, B. G. & Strauss, A. L. (1967). *The Discovery of Grounded Theory*. Aldine de Gruyter.

Hays, S. (1996). *The Cultural Contradictions of Motherhood*. Yale University Press.

Hochschild, A. R. (1979). Emotion work, feeling rules, and social structure. *American Journal of Sociology*, *85*(3), 551–575.

Holstein, A. & Gubrium, A. F. (2002). Active interviewing. In D. Weinberg (ed.), *Qualitative Research Methods* (pp. 112–126). Blackwell.

Jones, N. (2010). *Between Good and Ghetto: African American Girls and Inner-City Violence*. Rutgers University Press.

Lara-Millán, A. (2021). *Redistributing the Poor: Jails, Hospitals, and the Crisis of Law and Fiscal Austerity*. Oxford University Press.

Luker, K. (2008). *Salsa Dancing into the Social Sciences: Research in an Age of Info Glut*. Harvard University Press.

Manzo, L. C. & Brightbill, N. (2007). Toward a participatory ethics. In *Participatory Action Research Approaches and Methods* (pp. 59–66). Routledge.

Morris, A. (2015). *The Scholar Denied: W. E. B. Du Bois and the Birth of Modern Sociology*. University of California Press.

Oakley, A. (1981). Interviewing women: A contradiction in terms. In H. Roberts (ed.), *Doing Feminist Research* (pp. 30–61). Routledge and Kegan Paul.

Pascoe, C. J. (2011). *Dude, You're a Fag: Masculinity and Sexuality in High School*. University of California Press.

Ray, R. (2017). *The Making of a Teenage Service Class: Poverty and Mobility in an American City*. University of California Press.

Ray, R. (2021). Ethnographers' circle. Presentation at the Annual Meeting of the Pacific Sociological Association, March 19.

Small, M. L (2009). "How many cases do I need?": On science and the logic of case selection in field-based research. *Ethnography*, *10*(1): 5–38.

Small, M. L (2015). De-exoticizing ghetto poverty: On the ethics of representation in urban ethnography. *City & Community*, *14*(4), 352–358.

Sprague, J. (1997). Holy men and big guns: The can(n)on in social theory. *Gender & Society*, *11*(1), 88–107.

Stacey, J. (1988). Can there be a feminist ethnography? *Women's Studies International Forum*, *11*(1), 21–27.

Strauss, A. (1987). *Qualitative Analysis for Social Scientists*. Cambridge University Press. doi:10.1017/CBO9780511557842

Sweet, P. L. (2020). Who knows? Reflexivity in feminist standpoint theory and Bourdieu. *Gender & Society*, *34*(6), 922–950.

Todd, Z. (2016). An Indigenous feminist's take on the ontological turn: 'Ontology' is just another word for colonialism. *Journal of Historical Sociology*, 29(1), 4–22.

Tuck, E., McKenzie, M., & McCoy, K. (2014). Land education: Indigenous, post-colonial, and decolonizing perspectives on place and environmental education research. *Environmental Education Research*, *20*(1), 1–20. https://doi.org/10.1080/13504622.2013.877708

Qin, D. (2016). Positionality. In *The Wiley Blackwell Encyclopedia of Gender and Sexuality Studies*. https://doi.org/10.1002/9781118663219.wbegss619

Varpio, L., Ajjawi, R., Monrouxe, L. V., O'Brien, B. C., & Rees, C. E. (2017). Shedding the cobra effect: Problematising thematic emergence, triangulation, saturation and member checking. *Medical education*, *51*(1), 40–50.

Weiss, R. S. (1994). *Learning from Strangers: The Art and Method of Qualitative Interview Studies*. The Free Press.

Williams, C. L. (1991). Case studies and the sociology of gender. In J. Feagin, A. Orum, & G. Sjoberg (eds.), *A Case for the Case Study* (pp. 224–243). University of North Carolina Press.

PART IV

Statistical Approaches

21 Data Cleaning

Solveig A. Cunningham and Jonathan A. Muir

Abstract

High-quality data are necessary for drawing valid research conclusions, yet errors can occur during data collection and processing. These errors can compromise the validity and generalizability of findings. To achieve high data quality, one must approach data collection and management anticipating the errors that can occur and establishing procedures to address errors. This chapter presents best practices for data cleaning to minimize errors during data collection and to identify and address errors in the resulting data sets. Data cleaning begins during the early stages of study design, when data quality procedures are set in place. During data collection, the focus is on preventing errors. When entering, managing, and analyzing data, it is important to be vigilant in identifying and reconciling errors. During manuscript development, reporting, and presentation of results, all data cleaning steps taken should be documented and reported. With these steps, we can ensure the validity, reliability, and representative nature of the results of our research.

Keywords: Data Cleaning, Data Management, Quality Control, Quantitative Methods

Introduction

Poor data quality is the undoing of any research endeavor. A manuscript with poor-quality data is difficult, and indeed even unethical, to publish and can place the authors in disrepute (see Chapter 2 in this volume). The potential negative consequences of poor data are well recognized (Kaur & Datta, 2019; Sadiq et al., 2011), as interventions and programs of research built on poor-quality data can be futile or even harmful. Thus, ensuring high data quality is imperative to the research process and to informing policies and interventions (Batini et al., 2009).

It is important to approach data quality anticipating that there will be errors and that errors can occur at any stage of the research process (Dasu & Johnson, 2003; INDEPTH Network, 2002; Osborne, 2013; Van den Broeck et al., 2005). Errors are data points that provide information that is not valid, meaning not correct. Erroneous values may include information that was measured incorrectly, that was recorded incorrectly, or that was inadvertently shuffled or mis-assigned in a database. Some errors cannot be prevented; some may even never be identified. However, the more we understand the types of errors that can occur, when they are likely to occur, and what the implications are for our overall data, the better prepared we will be to prevent them, identify them when they do arise, and minimize their effects.

The concept of data cleaning refers to the steps taken to ensure high data quality (Oni et al., 2019; Van den Broeck et al., 2005); this is often defined according to characteristics such as accuracy, relevance, timeliness, completeness, and consistency (Batini & Scannapieca, 2006; Kaur & Datta, 2019; Redman, 2001). However, the term data cleaning itself is rather misleading. Unlike dirt on our hands or stains on our clothes, data are not improved by a good scrubbing. They are not soiled with unwanted particles that can be rinsed away, leaving an immaculate, valid data set. Rather, data are composed of elements – some of them correct, others erroneous. Some of these erroneous elements can be identified, but others cannot; some correct elements may be mistaken as being erroneous. The options available to identify and address errors differ at each stage in the data process (Batini et al., 2009; Van den Broeck et al., 2005). Importantly, rash and misinformed efforts at data cleaning can introduce new errors; this can happen by replacing correct values with incorrect values, by replacing incorrect values with other incorrect values, or by causing larger damage to the data set (e.g., by accidentally deleting, duplicating, or shuffling records). See Box 21.1 for a list of terms used throughout the chapter.

This chapter sets out the steps of data cleaning as a component integrated into each stage of the data process. During the design of a study, we employ a process-driven strategy (Batini et al., 2009, p. 5) to delineate the data quality procedures that should be carried out throughout the study; thus, the focus is on anticipating and preventing errors. Prevention of errors is the priority during data collection (Osborne, 2013; Van den Broeck et al., 2005). During the steps of data entry, management, and analysis, we employ a data-driven strategy (Batini et al., 2009), concentrating on steps to identify, reconcile, and resolve errors, or minimize their effects; these are steps that are most aligned with the standard perception of data cleaning. At the same time, we must maintain vigilance to prevent additional errors from being introduced. Data entry and management should be planned to overlap with data collection, as this overlap expands the opportunities to reconcile and correct errors (Van den Broeck et al., 2005). During manuscript development and presentation of reports and findings, the focus is on documenting and reporting all steps taken in data cleaning.

Anticipation of Errors

There are many considerations when developing a project that involves data collection. These include setting the research aims, convening diverse priorities and community and academic partnerships, drafting the study protocol, securing funding, and developing the data collection instruments (see Chapters 5 and 6 in this volume). This is the time to also begin "data cleaning." The goal of data cleaning at this stage is to anticipate errors. All studies will have some errors, but careful planning can prevent some, can reduce others, and can set in place the tools to address many.

Box 21.1 Key terms related to data cleaning

Back-translation: Materials initially not developed in the language of data collection are translated into the language(s) of study administration by a bilingual translator; the resulting draft is then translated back to the original language by another bilingual translator and discrepancies from the original are examined and reconciled (see Brislin, 1970; Brislin and Freimanis, 2001 for details).

Data cleaning: The processes of anticipating, preventing, identifying, reconciling, resolving, and reporting data errors (see Van den Broeck et al., 2005 for details).

Data-driven strategy: Improving data quality by direct modification of data values (see Batini et al., 2009 for details).

Data editing: Changing data points with incorrect values.

Data management: The acquisition, validation, protection, and processing of data; this involves data storage within a formal database (e.g., a relational database management system).

Data analysis: Examining, transforming, and/or modeling data to generate useful information.

Double data entry: Dual entry of every data point from interview forms, ideally by separate data clerks, into a pre-programmed database; this is then verified by a supervisor, who also calculates an error rate. The process is used to ensure data quality when transferring data from a paper-based questionnaire to a data management platform.

Error rate: A measure of data quality wherein the number of errors is divided by the total number of data points (for an example, see Database Error Rate, 2008).

Hard cut-offs: Benchmark values used to define upper and lower limits in a data range that identify impossible values requiring immediate diagnosis (see Van den Broeck et al., 2005 for details).

Impossible values: Data points that are not theoretically or biologically feasible or plausible.

Incorrect possible values: Data points that are plausible or possible but not factually correct.

Inliers: Data values that fall within the expected range, whether or not they are correct.

Legitimate missing values: An absence of data in circumstances where this lack of information is appropriate or expected (e.g., a question is not applicable for a specific respondent – see Osborne, 2013 for details).

Illegitimate missing values: An absence of data in circumstances where this lack of information is inappropriate or unexpected (e.g., due to refusal to respond by the participant or an accidentally skipped question or recording of response – see Osborne, 2013 for details).

Outliers: Data values that fall outside the expected range. Outliers may be impossible values or true extreme values and may be correct or incorrect (see Kaur and Datta, 2019 for details).

Process-driven strategy: An approach to improving data quality by optimizing the processes that generate or revise data (see Batini et al., 2009 for details). Examples of techniques used within a process-driven stratey include process control, which involves the implementation of check and control procedures when data are created, updated, or accessed, and process redesign, which involves eliminating causes of poor data quality and introducing new activities to the data-generating process to improve quality.

Error rate: A measure of data quality wherein the number of errors is divided by the total number of data points (for an example, see Database Error Rate, 2008).

Reliability: A component of data quality measurement in which the same value is obtained upon repeated measurement of the same phenomenon; it exemplifies consistency in measurement (see Koepsell and Weiss, 2014 for details).

Soft cut-offs: Benchmark values used to define upper and lower limits in a data range that identify suspect values requiring further screening (see Van den Broeck et al., 2005 for details).

Suspect values: Data points that are theoretically or biologically feasible, but fall beyond the expected range (see Van den Broeck et al., 2005 for details).

Validity: A data quality measurement in which the correct value is obtained for a given phenomenon (see Koepsell and Weiss, 2014 for details).

Research Protocol

At the outset of a new research program, the study team develops a study protocol that will serve as the guiding document for the project. In the protocol, the team specifies the purpose and objectives of the study, lays out the methodology that will be used to achieve each objective, and plans all aspects of data collection and analysis. Data quality is generally not a stated study objective, but it is an assumption that data will be of adequate quality for the objectives to be met. It is useful to build into the study protocol a section on "Methods to ensure data quality at each study stage." Alternatively, the team may decide to include a subsection in each section of the protocol describing the data-quality steps that will be taken at each stage of the study. Table 21.1 provides an example of the considerations that should be included in the protocol during study planning. Some teams preregister their study protocols before beginning study activities, to demonstrate *a priori* their approach to addressing the research questions; outlining beforehand the steps taken to ensure data quality can signal the rigor of the study to potential users of the data.

Workplan

The workplan should include sufficient time for data verification and management. A pre-test phase should be planned, during which the validity and reliability of the data collection instruments is enhanced. Additionally, a pilot test phase should be planned shortly before the implementation of the main data collection; this is the "dress rehearsal" for the study, implemented to verify that data collection and other aspects of the fielding process are operating as planned. Data entry and management should begin as part of the pilot test. In the workplan, data entry and cleaning should overlap with data collection. This overlap ensures that errors and potential

Table 21.1 *Study protocol considerations related to data cleaning*

Study stage	Data-quality threats	Prevention methods	Identification methods	Treatment methods
Study design	Unclear objectives Insufficient budget Lack of expertise Unclear, imprecise, or biased questions/measurements Long or difficult instrument	Validate data collection tools Hire experienced research staff Engage experts in the local context Provide training for research staff Predefine data cut-offs/restrictions Employ pretesting and pilot testing Translate/back-translate instrument	Plan overlap of data entry and cleaning with data collection	
Data collection/ data entry	Non-representative sampling Low response rate Poorly trained, un-cooperative, dishonest field staff Data-capture problems (hardware and software issues)	Use up-to-date sample frame Supervision of field teams Establish clear reporting system Daily debriefing to discuss issues Assign field teams to specific samples Refresher training for research staff Double data or validated data entry	Field-data checks Descriptive examination Visual-record examination Data error-rate estimation Fieldworker documentation	Recontact respondent(s) to address missing or impossible values
Data management	Database platform vulnerable to errors Programming or software problems	Use database management software Maintain locked original version of the data Supervision by senior analyst(s)	Descriptive examination of summary statistics and frequency distributions Visual-record examination	Perform all data cleaning using a working data file; securely store original data

Table 21.1 *(cont.)*

Study stage	Data-quality threats	Prevention methods	Identification methods	Treatment methods
	Introduced data errors Loss of data (deletion or scrambling)	Verify correct values prior to changing	Error-rate estimation Algorithm-based screening Graphical exploration of data Statistical outlier detection	Data interpolation or imputation for missing values Review data flow to inform correction of impossible values
Data analysis	Insufficient statistical skills Insufficient familiarity with data	Use formal analytical software (e.g., R) Use clear variable names Reference the data dictionary and interview guide when re-coding variables	Perform descriptive analysis of summary statistics and frequency distributions Graphical exploration of distributions Statistical outlier detection	Perform analytical data transformations on recoded variables Maintain unaltered original data set

problems are identified early and can be addressed. The workplan should set the development of the data entry program before the data collection instruments are piloted and should be tested during pilot testing. This is the case whether data will be entered in the field, directly into tablets or computers, or whether data will be collected on paper and subsequently entered into a database from the paper files.

The workplan should allocate sufficient time for data management and analysis during and following data collection. Typically, at least one month is needed, even for a small study. Data collected via computer or tablets requires a similar time frame for data management and analysis; for these studies, there is generally not an additional stage of data entry, and automated restrictions can be used to reduce erroneous data capture and reduce the need to do post hoc data cleaning. On the other hand, other errors are often introduced through data collection technologies, such as accidentally skipped modules, duplicate records, and incorrect values that are within the expected range of values. As such, the team should plan the time to check the data entry system and the data carefully, even when no paper forms are involved.

Budget

The budget should include resources for the data-quality steps. Most of the resources needed relate to personnel. The budget must be adequate to hire sufficiently qualified, and, if possible, experienced staff to conduct data collection, data entry, and data management. It must include time for training staff on the standards and methods to be applied and on the equipment that will be used. The costs of the equipment used for data collection, data entry, and data management must be calculated as well as needs for training staff in the use of the equipment and experts for troubleshooting of equipment. Teams may consider data-entry platforms that are free, such as EpiInfo for data entry and management and R for data analysis (Dean et al., 2011; R Core Team, 2013); others may opt for software that requires paid licenses, such as Microsoft Access or RedCap for data entry and management and Stata for data analysis (Harris et al., 2009; Harris et al., 2019; StataCorp, 2021). Some examples of options for data entry and management are provided in Table 21.2 (additional options are documented by Capterra; www.capterra.com). Paid programs do not necessarily guarantee a better experience; generally, the platform that is most familiar to the team members will be the one that will be easiest to use. Data collection and management equipment needs include paper and computers; some projects will use tablets for data collection; some also include scales, vials, and other tools for taking biometric direct measurements or refrigerators for storing samples. Some studies additionally need to budget for consultants to program and maintain the equipment.

Team

At least one of the core scientific team members should have prior experience with data collection and data management. A consultant can be added to advise on these methods if the expertise on the team is limited. Core scientific team members should

Table 21.2 *Common software platforms for data entry, data management, and data analysis*

Software platform	Common uses	Learning curve	Cost consideration	Pros	Cons
Alchemer (SurveyGizmo)	Data collection/entry	Easy	Paid	Internet-based surveys, large range of features for building surveys, intuitive user interface	Relatively expensive – especially for more powerful versions
Google Forms	Data collection/entry	Easy	Free	Integrates smoothly with other Google products/tools, simple interface with ride range of questions	Limited customization
ODK Collect	Data collection/entry	Moderate	Free	Facilitates electronic data capture, automated data entry, open source	Only for Android-based devices
Qualtrics	Data collection/entry	Moderate	Paid (Free options may be available)	Intuitive user interface, flexibility in creating survey questions, large suite of built-in features	Varied experience with customer support, relatively expensive for more powerful versions
EpiInfo	Data collection/entry Data management Data analysis	Easy	Free	Large suite of data collection, management, and analysis tools specifically designed for public health surveillance	Only for Windows-based applications
REDCap	Data collection/entry	Moderate	Free to non-profit organizations that join REDCap Consortium	Facilitates electronic data capture, secure data collection, HIPPA* compliant	Requires sufficient internal IT support; external IT contracting is not permissible Limited customization
Microsoft Access	Data entry Data management	Moderate	Paid	Large user community Integrates with Microsoft SQL Server	Requires coding skills similar to SQL Only for Windows-based platforms

Microsoft SQL Server	Data Management	Moderate	Paid	Data security, user-friendly interface	Requires SQL coding skills, relatively expensive – especially for more powerful versions
MySQL (Oracle)	Data Management	Moderate	Free	Large developer/user community, open source, compatible with multiple operating systems (Linux, MacOS, Windows)	Requires SQL coding skills
Python (PyCharm; other platforms available)	Data analysis	Difficult	Free	Open source, large developer community, high versatility with potential for customization	Requires programming language skills
R (R Studio; other platforms available)	Data analysis	Difficult	Free	Open source, large developer community, high versatility with potential for customization	Third-party package dependent for advanced capabilities, requires programming language skills
SAS	Data analysis	Moderate	Paid	Large range of "out of the box" functionality	Functional, but average quality graphics/visualizations, relatively expensive
SPSS	Data analysis	Easy	Paid	Beginner friendly, large range of "out of the box" functionality, user-friendly graphical interface	Limited potential for customization – particularly for more advanced analyses, low-quality graphics, relatively expensive
Stata	Data analysis	Moderate	Paid	Beginner friendly, large range of "out of the box" functionality, user-friendly graphical interface	Functional, but low-quality graphics, relatively expensive for more powerful versions

HIPPA: US Health Insurance Portability and Accountability Act of 1996.

also have experience with the software they select for data entry, management, and analysis. The field team should include sufficient data collection staff and supervisors to collect the amount of data needed given the amount of data and the duration of data collection. The number of staff needed is calculated by considering how many worker-days are required to achieve the calculated sample size, when accounting for the duration of administering each instrument, and the time needed to reach each sampled respondent. The data management team should begin work at the same time or shortly after the data collection team, as overlapping data collection and management provide opportunities to identify and correct errors in the field. Both the field team and data management team should be trained to understand the entire study, the necessity of data quality, their specific roles, and any equipment they must use. Both teams need a strong supervisory structure to ensure that procedures are followed systematically.

Formative Stages

What is the range of possible responses or values we should expect for a given survey question or measurement? What are the biologically plausible values for a given biomarker? This type of information is necessary to determine which data values are dubious or impossible and, therefore, require further examination and possibly correction. This information comes from engaging subject-matter experts at the early stages of a project, especially in developing and testing data collection instruments and setting up the database. In international studies, having local experts on the team provides contextual knowledge. Spending time to understand the study population also contributes to the team's ability to prevent and identify errors. For example, depending on the scope of the study, the team needs to establish a priori what the biologically possible and biologically plausible values are for a baby's weight at birth, for blood sugar levels (e.g., measured through a hemoglobin A1c test), for change in a child's reading scores between the beginning and completion of kindergarten, etc. Hard cut-offs must be established to disallow the entry of impossible values; hard cut-offs delineate which values are outside the possible range – and must never be captured – and which are within the possible range – and can be captured. Of course, just because a value is within the possible range does not mean that it is valid.

In addition to hard cut-offs, the team, including through expert consultations, also determines soft cut-offs. Soft cut-offs establish what values are possible to observe but are very rare or unlikely to occur. For example, it may be possible for a person to have been married 10 times, but it is unlikely. This process establishes what values need to be identified as extreme but plausible values; these values will be allowed to be recorded but will be flagged for additional investigation to determine whether or not they are correct.

One of the advantages of using devise-based data collection is that hard and soft cut-offs can be programmed before the beginning of data collection. These settings generate a warning to the person entering data if they are attempting to enter a value outside of the soft cut-offs and outside of the hard cut-offs; some programs make it impossible to enter a value beyond the hard cut-offs and do not allow data collection to proceed to the following item until a value in the expected range is entered. These steps can prevent data collection staff from accidentally recording incorrect values with a typo; they may prompt data collection staff to verify if they misheard or if a question was misunderstood. Special care must be taken with setting hard cut-off restrictions that do not allow out-of-range values to be entered: The team must be sure that the cut-offs do not inadvertently prevent values that are extreme but possible from being entered; doing so would entail that the data collection staff are compelled to enter false data to be able to move on to the following question. For this reason, it is generally better to rely on warnings rather than restrictions in setting up data-entry databases.

Prevention of Errors

Having anticipated possible sources of error at the study development stage, there is now a clear plan in place for preventing errors during the stages of fieldwork and data collection. It is important to monitor and to adapt the plans for error prevention as new challenges or data requirements are identified and new knowledge is gathered.

Data Collection Instruments

To prevent errors, data collection instruments must be carefully developed and informed by formative research and input from experts, as described above. Unless they are drawing heavily from instruments already validated in the same study population, a validation stage should be implemented. It is important to keep data collection instruments short while still meeting the study objectives; long, time-consuming, and demanding questionnaires deter sampled respondents from participating or lose their attention midway, with negative implications for generalizability and validity. Once drafted, instruments should be meticulously improved through multiple iterative versions during the pre-test stage. The instrument must be pre-tested repeatedly and then pilot-tested in the study population.

Training

The field team should be well trained to understand what types of errors may threaten data quality, how and when such errors occur, and what can be done to prevent them. Specific areas for training should include abiding by the sample selection procedures from the sampling frame, collecting respondent-reported and directly measured data correctly, interacting with respondents as directed, not influencing respondents' answers (even unintentionally), interpreting and recording the information shared by respondents correctly, not falsifying data, handling complex and atypical responses from participants, and reporting procedures to follow if assistance is needed.

Supervision

A strong supervisory structure helps to reduce errors by ensuring that data collection is proceeding following the protocol and by speeding up the identification and resolution of any errors that do occur. Supervision is conducted on a daily basis by field supervisors, who should spend about half of their time monitoring their team's data quality, in addition to also conducting data collection. Study senior leadership should also be engaged in regular supervision, as their presence in the field keeps staff morale high and signals to the field team, and to participants, the importance of the study; this multi-pronged supervision promotes high response rates and accurate data.

The supervisor should schedule unannounced drop-ins to ensure that data collection is proceeding according to the protocol. An additional measure for checking for errors is to have supervisors re-interview 10% of participants and re-collect a subset of the questions or measurements; it is expected that most of the answers will be very close to those generated by the initial interview. If they are substantially different, or if the respondent says that they were never approached by an interviewer, then the supervisor should investigate whether there was cheating or miscommunication on the part of the interviewer.

Translation

Non-coverage error occurs when a segment of the population does not have a chance of participating in a study. This can happen if a person does not know the language used in the data collection or does not know it sufficiently well to be willing to participate. The study can only generalize to those people who have a chance of participating in the study, so data must be collected in each language of the population to which the study will generalize. If the study is initially not developed in the language in which the data will be collected, after the initial pretests, materials should be translated and then back-translated by a different bilingual person (Brislin, 1970; Muir et al., 2018, 2019, 2020b). This process ensures that the phrasing of questions and response options being collected is true to the study design. Having interviewers translate questions during the interview is not recommended, as this reduces precision and introduces inconsistencies across respondents. In some cases, interpreters are hired to translate between interviewers and respondents; this approach has the same shortcomings. The data collection team should include at least one person who will be able to communicate in each of the languages of the study. Field pre-testing and pilot-testing should be conducted in all languages.

Data Restrictions and Cut-offs

Predefining – through formative research, reviews of the literature, and consultation with experts – what are possible, improbable but possible, and impossible values, allows the team to establish the range of acceptable values for each indicator in the data collection instrument. This instrument should have specific instructions about what values may be entered. In device-based data collection, the research team can program hard cut-off restrictions, so that it is not possible to enter what are considered impossible or invalid data (e.g., a pregnant male). Thus, it will be harder for the field team to accidentally enter a value (e.g., an age of 200 instead of 20 or an incorrect sex).

However, the team should be careful with the use of hard cut-offs, as they can also introduce errors. For example, imagine a scenario where the age range established by the team for respondents to be asked about pregnancy is ages 15 to 45. What should the interviewer do if she encounters a 14-year-old pregnant woman? In paper-based data collection, she might record the woman's age and pregnancy and make a note in the margin or in her field notes explaining that she confirmed that the woman was 14 years old and yet pregnant; this can then be discussed with supervisors and the research team at the end of the day, and a decision on how such situations should be handled can be reached collectively. However, on device-based data collection, if hard cut-offs are set that only allow data in the expected range to be entered, then the interviewer is forced to enter incorrect data – she must either incorrectly report that the woman is not pregnant or incorrectly record the woman's age as 15. In self-administered questionnaires, it is the respondent who must decide how to change the information to be able to provide a response; in some cases, a respondent to a self-administered survey might stop participating in the study in frustration. Either scenario introduces errors that will likely not be detected because no impossible or improbable values are generated in the process, so the team members do not even realize that there is an error.

Data Entry

An important source of error is data entry. This is the case whether data entry is done in the field, through devise-based data collection, directly into a database, or in the office from paper-and-pencil interview files. For example, a data-entry clerk may accidentally enter one 0 too many or too few on a numeric value, such that a person with 10 years of school is recorded as having just 1 (or 100) year(s) of school. Similarly, he might select response option B, which the respondent had indicated for question 29, but incorrectly record it for question 30.

With paper-and-pencil interviews, the possibilities for data cleaning are more numerous. One option for identifying and reconciling errors is double data entry (Barchard & Pace, 2011; Cummings & Masten, 1994; Kawado et al., 2003). With this method, two data-entry clerks independently enter data into the pre-programmed database, and the discrepancies between the two are then checked by a supervisor (e.g., by calculating an error rate manually or through the use of formal statistical analysis). If full data entry is too costly, at least 25% randomly selected records should be double-entered. If errors are found, additional re-entry should be considered.

Some researchers have argued that the benefits of full double data entry may not offset the expense and time commitment required (Atkinson, 2012; Day et al., 1998; King & Lashley, 2000; Reynolds-Haertle & McBride, 1992). Day et al. (1998) suggest that because the data errors most likely to impact statistical analyses are detectable using simple range checks or other exploratory data analysis techniques, double data entry may not be necessary.

An additional method involves visual verification of the data to source documents on a record-by-record basis, sometimes referred to as visual-record verification (King & Lashley, 2000). This strategy is often implemented within a continuous sampling plan wherein the first 10 records of the data set created from single data entry are compared to their corresponding source document. If these records are correct, every 10th record is then checked until an incorrect entry is identified. If an incorrect entry is identified, the error is corrected, and the entered data are then fully checked until 10 accurate records are found. Thereafter, the process returns to checking every 10th record. This method is easy to use and not very costly or time-consuming. However, others have found that visual assessments miss more errors than double-entry techniques (Barchard & Pace, 2011).

Device-based data collection relies on the accuracy of the field staff. Even field teams doing device-based data collection should carry with them a few blank paper interviews, in case the device runs out of battery, freezes, or is otherwise unusable; the paper-and-pencil interview forms will allow data collection to proceed with the sampled participants, and the data can be entered later. Regardless of the mode of data collection, but especially when using device-based data collection, field staff should carry a notebook where they record any discrepancies that occur in data collection and data entry. For example, an interviewer might accidentally record an incorrect response, but, once it has been entered, the device does not allow her to go back and make a correction. By the end of the day, she will have forgotten the details; however, if she is able to make the correction or document the problem on paper, she can review it with the supervisor at the end of the day.

Data Cleaning and Management

Whether using double or single data entry, field-based, or office-based data entry, a supervisor or analyst needs to check for errors in the database. This involves checking each variable's range by tabulating the values at the mean, minimum, maximum, and the 75th and 99th percentiles; the skewness, kurtosis, and number of missing values should also be reviewed. Descriptive tables with variable frequency distributions and cross-tabulations and histograms are useful for visual inspection. Data management software can be programmed to automatically flag impossible values and discrepancies for manual review. Evaluation criteria may also include the calculation of an error detection rate specific to data entry; these calculations can be performed manually or via formal statistical analysis. Prior studies indicate data-entry error rates may range from less than 1% to 27% (Atkinson, 2012). Acceptable levels of overall error should be decided at the study outset of the study. For instance, *The Encyclopedia of Public Health* recommends that overall error rate standards are set below 1%; a common threshold for acceptable error is 0.1% (Database Error Rate, 2008).

The database for data entry and management should be set up in data management software (see examples in Table 21.1) and not in spreadsheets. If a spreadsheet is used, for example in Excel, the entire data set can be compromised by accidentally reshuffling the observations. For example, an analyst might accidentally re-sort a data set in a spreadsheet so that the values are attributed to the wrong respondent;

they might also accidentally delete or replicate a line, thus losing or duplicating an observation, or delete a column, thus deleting a variable.

The process of variable coding and labeling and of identifying missing and erroneous values is intended to improve data quality. However, these steps themselves can generate errors. For example, an analyst reviewing data on body weight may find that a participant is listed as weighing 310 kg; she decides that this is not possible, and that the data-entry clerk must have accidentally entered 310 when the correct answer was 130. Unless this suspicion is confirmed, the data should not be changed. To prevent new errors from occurring during data cleaning and management, it is imperative to preserve an original data set with no values altered – even values that are likely incorrect. The database should be locked at the completion of data entry. Any changes should be made into a copy of the database. Even in this "working version" of the database, changes should only be made when there is concrete evidence that a correct value has been identified. Any changes made in the working database should be automatically stamped with the identification number of the person who made the change and the date. The person making the change should also record, in a dedicated notebook or an electronic file, the changes made and the basis justifying these changes.

Data Analysis

Data analysis should begin with data checks and exploratory descriptive reviews of the data. This step is important even when using a data set that has already been reviewed and used by others. Before beginning to populate descriptive tables or run analytical models, the research team must first review each variable they are considering using. Similar to the steps carried out by the data management teams, analysts should tabulate, for each variable, the values at the mean, median, minimum, maximum, and percentiles of the variable's distribution, skewness, kurtosis, and number of missing values. This phase of understanding the data set will make it clear whether a variable has many missing, extreme, or unexpected values; it will also inform the researcher whether there is sufficient variation in the values to conduct regression analysis and whether the distribution of values is normal or is in line with expectations (see Chapter 22 in this volume).

Based on this information, the researcher will determine whether a variable is usable, as variables with extensive problematic or missing values, and variables with insufficient variation (e.g., 90% of responses with the same value) may need to be excluded from analyses. If the variable is usable, additional steps may be needed before analysis. Values indicating that the respondent did not know or refused to provide an answer are generally coded as missing for analysis. If a question or measurement was not applicable to a respondent and, therefore, was not collected from that respondent, this data point will also need to be coded as missing for analysis. If some of the values have a very small cell size (i.e., very few respondents gave a specific response), the researcher may decide to re-code the variable by combining several small cells; this is only done in situations where combining response categories is conceptually appropriate.

The data analysis stage is not exempt from the introduction of new errors. One example is that a researcher may mislabel variables – instead of naming a variable “number of years of completed education,” he might label it as “years working.” There are several steps that can prevent such problems. One solution is to use clear variable names in the database, including the question number from the instrument. When recoding variables, the researcher should have the interview guide and data dictionary close by and refer to the original question for each variable.

Another common confusion is to mistake missing values for true values. For example, on a variable measuring household income, the missing code for a response of “don’t know” might be pre-coded to be 999999. Using such missing data codes is standard procedure, but an unfamiliar researcher might interpret them as true values (e.g., as an income of $999,999). To avoid such situations, before beginning to populate tables or run analytical models, we must first review each variable, as described above.

Identification of Errors

Erroneous values appear, broadly speaking, in three categories: missing values, impossible values, and incorrect possible values. Researchers should be prepared to identify these at each point in the data process. Suspect and problematic values should be flagged for analysis. The earlier they are identified, the more can be done to fix them and to prevent other similar errors from occurring. If supervisors identify errors through random re-interviews or when reviewing data at the end of a day of data collection, the team can recontact participants right away and verify or correct the information. When errors are found during data management and analysis after data collection is complete, it is generally not possible to correct the data.

Missing Values

Missing values are data points that have no information (McKnight et al., 2007; Osborne, 2013; Van den Broeck et al., 2005). The presence of missing values can negatively impact scientific inquiry by affecting the reliability and validity of data. Specifically, missing values may directly affect construct validity – the extent to which a given measure captures the information or construct (variable) we intend to observe (McKnight et al., 2007, p. 20). Missing values may also indirectly affect internal validity, or a researcher’s ability to assert that a given factor affected an outcome of interest, by contributing to biases and other threats to validity within the data (McKnight et al., 2007). Values can be missing due to study design, participant characteristics, measurement characteristics, data management, and chance (see Table 21.3. for details).

A useful way to communication about the data is to use classifications of missing values according to the anticipated reason(s) for why the data are missing (i.e., the

Table 21.3 *Aspects of a study that may lead to missing data (see McKnight et al., 2007 for further details)*

	Contributing factors	Potential solutions
Study design	Number of measurement occasions Timing of data collection Number of variables Assignment of participants (if applicable)	Reduce response burden: limit the frequency of measurement limit the duration of measurement
Characteristics of target population and sample	Identification and recruitment of target group Perceptions about the study topic Socio-demographic characteristics	Piloting survey instruments Community sensitization and engagement Providing participation incentives Providing confidentiality assurance Increasing ease of participation
Data collection methods	Human or equipment error Observational error Interview/survey setting and duration Interviewer/enumerator characteristics Participant/respondent characteristics	Detailed research protocols Training of research personnel Random checks and evaluations Hire experienced data collectors
Instrument characteristics	Length of an instrument Content of an instrument Layout and format of an instrument	High-quality printing Large font sizes to facilitate reading Avoid overcrowding text Accurate alignment Editing for omissions and misspelling
Data entry	Tedious, dull task Unmotivated or poorly trained personnel	Providing incentives and feedback Data-entry validation

mechanisms of missingness; McKnight et al., 2007, pp. 40–41; Osborne, 2013, p. 109). Missing values can be broadly characterized as either *legitimate* or *illegitimate* (Osborne, 2013). Legitimate missing values are instances where the absence of information is appropriate. This is the case when a question does not apply to a certain participant, resulting in a legitimate skip of that item for that respondent. Illegitimate missing values occur when a respondent does not provide a response. This can happen if the respondent does not know the answer or does not wish to share the information (Osborne, 2013). Illegitimate missing values can also occur as an unintended consequence of the data cleaning process in which a researcher incorrectly deletes values for a given data point (McKnight et al., 2007).

Another commonly used classification scheme for missing data was proposed by Rubin (1976); it describes missing values as missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). With MCAR, there is no systematic relationship underlying how values are either observed or missing. In contrast, with MAR, there is a relationship with how values are observed but not with how they are missing. Finally, with MNAR, an underlying relationship exists with regards to how values are missing and may also exist with regards to how they are observed. These classifications are related to the potential bias that missing values may generate in statistical analyses. For example, the impact of MCAR is ignorable from a modeling standpoint given the lack of an underlying relationship – the missing values are randomly distributed across all observation, which may limit analytical power, but should not bias results (Osborne, 2013). In contrast, in the case of MNAR, there is a systematic relationship underlying how values are either observed or missing, entailing that they can generate biased results (McKnight et al., 2007; Osborne, 2013).

Some missing values can be avoided through carefully designed data collection instruments. The *identification* stage of data cleaning involves distinguishing these from other missing values that are more problematic – where a data point is missing because the respondent, the interviewer, or the data-entry clerk (depending on the mode of data collection) did not enter a value or even accidentally skipped an entire page or module.

Impossible Values

To identify impossible values, the team must have subject knowledge or must consult with experts who can provide this information. It is important to define before data collection begins what are impossible values, as described above. For device-based data collection, the team may program parameters so that impossible values cannot be entered. These restrictions must be set very carefully, ensuring that true extreme values are not accidentally excluded, as that would lead to biases and incorrect values being introduced into the data. For example, for a study on breastfeeding, the team may decide based on personal experience that children can only be breastfed up to the age of two years and so values of duration of breastfeeding above 24 months are not allowed to be entered. However, in reality, many children are breastfed past the age of two years. In such circumstances, the respondent or interviewer would be left to decide between two incorrect approaches; he can record that a child is two years old, with this being the closest allowable value to the true response, despite the child being four; alternatively, he can record the response option as “don’t know.” Either approach would lead to incorrect values entering the data.

Some impossible values can only be identified in cross-tabulations with other variables. For example, a woman who has had 12 children is certainly a possible value. However, if we tabulate age with number of children and find that the woman is 15 years old, then we see that 12 children would be an impossible value.

Incorrect Possible Values

Incorrect possible values are the most difficult to identify and, therefore, to address, as there is nothing to indicate a problem exists when examining these data. Our best chance for identifying incorrect values that are within the expected range for a variable is through data checks while the team is still in the field. For example, field teams and supervisors can perform random checks to identify errors and ascertain what information should be entered. When suspect values are identified during data checks, the team can contact respondents to confirm whether the value is correct. Another possibility is examining other variables. For example, a household may report that they get their water from a borehole; however, when examining the Global Positioning System location of the homestead, the team sees that there is no borehole within 10 km of the homestead – this indicates that there may be an erroneous data point. There are additional opportunities to identify improbable possible values in longitudinal data, where information is collected repeatedly over time from the same respondents. In this case, we can examine the most recent data in comparison with earlier data. For example, it is very reasonable that a 20-year-old is 1.63 meters tall. However, if data collected one year earlier indicate that the respondent was 1.8 meters, we have indication of a data error.

To identify errors introduced later in the data process, we can examine a random sample of paper surveys with the originally entered data and compare them with the data in the database. We can also examine the data-entry data set in comparison with the analysis data set for any errors that may have been introduced.

Reconciling Errors

Steps

When people think about data cleaning, most commonly the focus is on correcting errors. Just like finding a spot on a shirt or crumbs on a table, the researchers find erroneous data, scrub them away, and replace them with new, correct values. It is best to eliminate this vision of the process from our goals. What is and is not a correct data point is often difficult to distinguish; replacing incorrect data points with correct ones is even more difficult. Indeed, a zealous and inexperienced researcher can introduce at least as many errors as s/he resolves. Therefore, any deletion and replacement of values must always involve the following steps:

- Never make a change to the original data set, even if you feel quite certain that the change would be a correction.
- Make any changes in an analysis "working" data set – not the original database.
- Any change needs to generate an ID number of the person making the change and the date when the change was made.
- A log should be maintained, where researchers track any changes that were made and the reason for each change. This log can be maintained in a notebook or as an open-ended variable in the data set. It is especially helpful to have these notes integrated into a command file in the analysis software.
- Consider having all changes reviewed and confirmed by a second, preferably more experienced researcher.
- Consider maintaining two copies of each variable – one including revised values and the other maintaining original values (e.g., "age" and "age_r"). This approach allows the team to conduct analyses with and without the revised variables to test the effects of the erroneous data on the results.

Handling Missing Data

Missing data is the easiest data problem to identify. There are two questions to resolve. The first, for any missing value, is whether there is a way to replace it with a correct value. In most cases, this is not possible. If the team is still in the field, they can attempt to recontact the participant to try to elicit a response. If there are paper interview files or electronic files from earlier stages in the data flow, the team can check whether a value was reported by the respondent but was subsequently lost. In longitudinal data collection, if there are data on the respondent from a previous round of data collection, the team can consider putting in the value from the previous round as the best estimate for the current round. This strategy is an example of single imputation, sometimes referred to as the "last value carried forward" (McKnight et al., 2007, p. 174). If there are data from multiple previous rounds of data collection or from previous and subsequent rounds, the team can consider whether linear interpolation between the data rounds would be appropriate for fitting in a value to replace the missing value. Other imputation methods may be available.

A second consideration is whether all missing data are the same or whether there is information in some types of missing values. Imagine, for example, a questionnaire that asks "How many weeks pregnant are you?" Or "Have you ever eaten a meat substitute?" If a respondent indicates, "I don't know," the research team might combine this response with refusals and other types of non-response; alternatively, the team may decide that the respondent saying that she does not know is itself meaningful. They could take the position that this response indicates that there are implications that could be addressed through interventions or through information campaigns. If so, they may decide to use "don't know" as a response category for that variable. This decision is specific to the research and will be made accordingly by the research team. The method for handing missing data called "dummy-variable adjustment" can be used to determine whether "don't know" or other missing values are systematically associated with the outcome variables under study, and this information can contribute to deciding how to treat these situations. In the creation of data collection instruments, it is important to ensure that responses of "don't know" can be recorded separately from refusals and other response options; otherwise, these steps are not possible.

Handling Impossible and Suspect Values

Having established decisions about which values are considered impossible, identification of this subset of erroneous values is straightforward. As with missing values, the same efforts can be made to replace the impossible values with a correct one, either by recontacting the respondent or by finding initial correct values earlier in the data flow that may have been erroneously entered or spoiled.

Impossible values should not be left unchanged. If the correct value cannot be ascertained based on data acquired from the respondents, the value should be recoded as missing. The missing categories should be labeled to distinguish these previously recorded impossible values from other reasons for missing values. Once labeled as missing, the value can be considered for imputation or any other method for handling missing values selected by the team.

Incorrect possible values are the most difficult errors to identify; the team should be very cautious in changing values. Suspect values should be investigated using the methods just described, and if an error is found and a correct value ascertained, a change can be made. No change should be made without clear evidence directly collected from the respondent. A random sample of non-suspect values should also be checked, and if errors are found, further investigation should be conducted in additional observations.

Identifying suspect values is time-consuming. It is important to plan for and use that time, rather than moving directly to data analyses. As a first step, at the end of each day, the field supervisors and researchers should inspect the data collected that day. This can be done with a quick review of each interviewer's completed surveys from that day to identify the following: are any items blank; are any values unexpected or inconsistent; do any items have markings or notes taken by the interviewer. A quick discussion with the data collection team to ask them about difficulties or unexpected situations during the day is enlightening. Thus, many errors that are found can be reconciled expeditiously.

Once the data are in a database, analysis can begin by scrolling through the data to informally pick up errors. The spreadsheet should already have pre-programmed flags in place to automatically identify impossible and unlikely values. These flags draw the analysts' attention to problematic values to inspect. As the analyst become increasingly familiar with the data and possible issues, they may add more flags.

Using Descriptive Statistics for Data Cleaning

Next, the researchers can begin systematic descriptive statistics. First will be univariate analyses – for each value, generate the mean, median, minimum, and maximum values, and the number of missing values of each type for each variable. These can be automatically generated for an entire database and can be inspected for anomalies. For continuous variables, especially, histograms are helpful to visually identify unexpected distributions and outliers.

Bivariate distributions can be inspected for unexpected associations. A correlation matrix of variables can allow for quick inspection of the magnitude of co-variation and whether the association is in the expected direction. Cross-tabulations indicate the mean value of categorical and bivariate variables relative to each other – again, allowing the researcher to identify unexpected associations. These methods will generally not identify specific data points that are erroneous, nor provide resolutions for errors; rather, they indicate that, in the data set, there is an unexpected trend, possibly resulting from multiple problems.

Reporting Errors

As part of the diligent process of identifying and resolving problematic values, careful documentation must be maintained on the amount of missing, impossible, and suspect values. For missing values, the type of missing value (e.g., refusals and unknowns) should be documented. The number of observations each of these affects in the entire data set should be documented in reports and publications. The methods that were used to deal with these values for analysis should also be reported. Reports and publications should show results with and without missing values and explain how observations with missing data are different from those with complete data. Results with and without reconciled values should also be reviewed and, if different, these differences should be examined statistically and reported.

As part of the creation of an analysis data set, a data dictionary and codebook must be created. These documents describe the aims of the data collection, what questions were asked, what measurements were taken, the methods used, and the creation of any composite variables made up of multiple questions or measures (e.g., the body mass index, which is calculated based on variables of height and weight). The codebook also indicates the distribution of values across categories and the range of values and codes for each category, including missing values. Skip patterns and decisions about what constitutes impossible values are also included. Components of the field manual, or the entire manual, can be included in the documentation so that a data user will know what steps, rules, and procedures were in place.

As a potential final step in reporting that fosters transparency, it is becoming more common to prepare summary reports of cleaned data and have these reports published in journals such as *Data in Brief* (e.g., see Cope et al., 2020; Muir et al., 2020a). An expectation associated with publication of these summary reports is that clean, de-identified data are made publicly available to bolster the rigor of scientific inquiry through encouraging and facilitating replication of analytical research.

Conclusion

Erroneous and missing values are a reality of all data sets. A goal in data collection, data management, and data analysis is to minimize the number of errors and, to the extent to which some errors are not preventable, to reduce their impact on research findings. By expecting that there will be errors, we can develop systems to reduce their frequency, to identify errors when they do occur, and to resolve these. In this chapter, we have provided an overview of some of the methods available to researchers to prevent, identify, and resolve errors. We cannot rely on short, last-minute efforts to clean data, at the beginning of data analysis. To generate high-quality data requires anticipating errors at the planning stages of data collection, preventing errors during data collection and management, identifying errors as early as possible, and reconciling them only when a resolution is clearly verified.

This thorough approach has implications for project timelines and budgets, as error prevention and identification can be time-consuming and requires skill. At the same time, these steps, planned for and conducted during data collection and management, will save time during data analysis; they improve data quality by allowing us to correct problems. The steps we take in data cleaning need to be documented and reported so all data users will have a clear assessment of the quality of the data. The number of errors, sources of errors, and whether and how they were reconciled should all be reported.

Taken together, the steps of data cleaning at all stages of data collection and management reduce the number of errors that occur and indicate to the research team which data points need to be reconciled and how.

References

Atkinson, I. (2012). Accuracy of data transfer: Double data entry and estimating levels of error. *Journal of Clinical Nursing*, *21*, 2730–2735.

Barchard, K. A. & Pace, L. A. (2011). Preventing human error: The impact of data entry methods on data accuracy and statistical results. *Computers in Human Behavior*, *27*(5), 1834–1839.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys (CSUR)*, *41*(3), 1–52.

Batini, C. S. M. & Scannapieca, M. (2006). *Data Quality: Concepts, Methodologies and Techniques*. Springer.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology*, *1*(3), 185–216.

Brislin, R. W. & Freimanis, C. (2001). Back-translation. In D. E. Pollard (ed.), *An Encyclopaedia of Translation: Chinese–English, English–Chinese* (pp. 22–41). Chinese University Press.

Cope, M. R., Slack, T., Blanchard, T. C., Lee, M. R., & Jackson, J. E. (2020). The Louisiana community oil spill survey (COSS) dataset. *Data in Brief*, *30*, 105390.

Cummings, J. & Masten, J. (1994). Customized dual data entry for computerized data analysis. *Quality Assurance (San Diego, California)*, *3*(3), 300–303.

Dasu, T. & Johnson, T. (2003). *Exploratory Data Mining and Data Cleaning, Volume 479*. John Wiley & Sons.

Database Error Rate (2008). Database error rate. In W. Kirch (ed.), *Encyclopedia of Public Health* (pp. 196–196). Springer Netherlands. https://doi.org/10.1007/978-1-4020-5614-7_667

Day, S., Fayers, P., & Harvey, D. (1998). Double data entry: What value, what price? *Controlled Clinical Trials*, *19*(1), 15–24.

Dean, A., Arner, T., Sunki, G., et al. (2011). Epi Info™, a database and statistics program for public health professionals. CDC, Atlanta, GA.

Harris, P. A., Taylor, R., Minor, B. L., et al. (2019). The REDCap consortium: Building an international community of software platform partners. *Journal of Biomedical Informatics*, *95*, 103208.

Harris, P. A., Taylor, R., Thielke, R., et al. (2009). Research electronic data capture (REDCap): A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, *42*(2), 377–381.

INDEPTH Network (2002). *Population and Health in Developing Countries: Volume 1; Population, Health, and Survival at INDEPTH Sites*. IDRC.

Kaur, A. & Datta, A. (2019). Detecting and ranking outliers in high-dimensional data. *International Journal of Advances in Engineering Sciences and Applied Mathematics*, *11*(1), 75–87.

Kawado, M., Hinotsu, S., Matsuyama, Y., et al. (2003). A comparison of error detection rates between the reading aloud method and the double data entry method. *Controlled Clinical Trials*, *24*(5), 560–569.

King, D. W. & Lashley, R. (2000). A quantifiable alternative to double data entry. *Controlled Clinical Trials*, *21*(2), 94–102.

Koepsell, T. D. & Weiss, N. S. (2014). *Epidemiologic Methods: Studying the Occurrence of Illness*. Oxford University Press.

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing Data: A Gentle Introduction*. Guilford Press.

Muir, J. A., Braudt, D. B., Swindle, J., Flaherty, J., & Brown, R. B. (2018). Cultural antecedents to community: An evaluation of community experience in the United States, Thailand, and Vietnam. *City & Community*, *17*(2), 485–503.

Muir, J. A., Cope, M. R., Angeningsih, L. R., Jackson, J. E., & Brown, R. B. (2019). Migration and mental health in the aftermath of disaster: Evidence from Mt. Merapi, Indonesia. *International Journal of Environmental Research and Public Health*, *16*(15), 2726.

Muir, J. A., Cope, M. R., Angeningsih, L. R., & Brown, R. B. (2020a). Community recovery after a natural disaster: Core data from a survey of communities affected by the 2010 Mt. Merapi eruptions in Central Java, Indonesia. *Data in Brief*, *32*, 106040.

Muir, J. A., Cope, M. R., Angeningsih, L. R., & Jackson, J. E. (2020b). To move home or move on? Investigating the impact of recovery aid on migration status as a potential tool for disaster risk reduction in the aftermath of volcanic eruptions in Merapi, Indonesia. *International Journal of Disaster Risk Reduction*, *46*, 101478.

Oni, S., Chen, Z., Hoban, S., & Jademi, O. (2019). A comparative study of data cleaning tools. *International Journal of Data Warehousing and Mining (IJDWM)*, *15*(4), 48–65.

Osborne, J. W. (2013). *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data*. SAGE Publications.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. Available at: www.r-project.org.

Redman, T. C. (2001). *Data Quality: The Field Guide*. Digital Press.

Reynolds-Haertle, R. A. & McBride, R. (1992). Single vs. double data entry in CAST. *Controlled Clinical Trials*, *13*(6), 487–494.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Sadiq, S., Yeganeh, N. K., & Indulska, M. (2011). 20 years of data quality research: Themes, trends and synergies. Proceedings of the Twenty-Second Australasian Database Conference, Perth, January 17–20, Volume 115,

StataCorp (2021). Stata statistical software: Release 17. StataCorp LLC.

Van den Broeck, J., Argeseanu Cunningham, S., Eeckels, R., & Herbst, K. (2005). Data cleaning: detecting, diagnosing, and editing data abnormalities. *PLoS Medicine*, *2*(10), e267.

22 Descriptive and Inferential Statistics

Martha S. Zlokovich, Daniel P. Corts, and Mary Moussa Rogers

Abstract

What are statistics and why do we need them? This chapter introduces descriptive statistics and then creates a bridge from describing data concisely to answering questions using hypothesis testing and inferential statistics. The chapter leads the reader to an understanding of how descriptive statistics summarize and communicate meaning, based on data, and how they underpin inferential statistics. Research study examples, figures, and tables throughout the chapter explain the topics addressed by applying the ideas discussed. The chapter begins with the basics of descriptive statistics – normal distributions, options for displaying frequencies, measures of central tendency and variability, and correlations. The transition to inferential statistics covers standardization and the *z*-score, sampling, confidence intervals, and basics of hypothesis testing including Type I and II errors. We then introduce inferential statistics using three methods – *t*-tests, one-way analysis of variance (ANOVA), and chi-square tests.

Keywords: Descriptive, Scales of Measurement, Frequencies, Central Tendency, Inferential, Analysis of Variance, *t*-Test, Chi-Square Test

Introduction

This chapter introduces descriptive and inferential statistics, leading the reader to an understanding of how they summarize and communicate meaning, based on data, and how they underpin inferential statistics. The descriptive statistics introduction then leads from describing data concisely, to answering questions using hypothesis testing, to inferential statistics. Illustrations of descriptive and inferential statistics throughout the chapter are based on research by Smith (undergraduate student at the time) and Smith's faculty sponsor, Ransford (Smith & Ransford, 1999), regarding the relationship between disordered eating and conformity among sorority member and non-sorority member college students.

The chapter begins by explaining the basics of descriptive statistics, including scales of measurement, options for displaying frequencies, normal distributions, measures of central tendency, and measures of variability. Next, the chapter draws connections between descriptive statistics and inferential statistics, noting how assumptions related to normal distributions are essential. The transition topics covered include standardization and the *z*-score, confidence intervals, correlations, hypothesis testing, Type I and Type II errors, and hypothesis testing using the *z*-test. The final section introduces inferential statistics by describing *t*-tests, one-way analysis of variance (ANOVA), and chi-square tests. These inferential methods are introduced in the context of extending the five steps for null hypothesis significance tests (NHST) for *z*-tests to additional methods of data collection.

Descriptive Statistics

Descriptive statistics provide readers with an overview of the variables explored in a research project and summarize the most basic research outcomes. They most often appear in the results section or in the methods section of a research paper; in the latter, descriptive statistics appear in descriptions of participants' demographic characteristics (e.g., age, gender, and ethnicity). In some cases, descriptive statistics are the only option for presenting a summary of the findings in the results section; in other cases, they set the scene for the hypothesis testing needed to infer causation among variables that will follow. Descriptive statistics generally are not used to test the primary hypothesis in a study because they do not address causation. Nonetheless, it is important for researchers to thoughtfully present descriptive statistics in a way that makes the numbers easier to understand, and that correctly conveys the facts revealed by the data.

Scales of Measurement

When evaluating a concept, researchers must decide exactly what to measure. This begins by defining concepts of interest in terms of observable, measurable variables. Variables are observed in a variety of ways that are categorized according to scales of measurement, which each have characteristics that are key to properly evaluating a concept and the types of statistical analyses that can be conducted with them. There are generally four types of scales, as detailed below.

Nominal Measurement

Nominal measurement is defined by qualitative differences. The meaning of the observations determines differences in nominal variables. Simply put, nominal measurement is based on categorizing or classifying – mother, father, grandmother, and grandfather are categories of the variable *relative*. Numbers can be attached to categories in nominal scales (e.g., mother = 1, father = 2, grandmother = 3, grandfather = 4) to allow statistical analysis but must be used with caution so that the statistical analysis does not presume mathematical differences (e.g., $1 < 2$). Nominal measures typically have to do with demographic categories (e.g., race/ethnicity and gender identity) or research design groups (e.g., control group and experimental group). As another example, the study we present in this chapter compares members of a sorority (a membership-based social group associated with a university) to students who are not members – two distinct, nominal categories. Nominal variables cannot be compared in terms of greater than or lesser than. Extending our example, mother is not numerically greater than or less than father, grandmother, or grandfather.

Ordinal Measurement

Ordinal measurement might be thought of as nominal plus – the plus being that one value is greater or less than another. Ordinal measurement is defined by ranking observations. Ranking, however, does not indicate the variation or any numerical difference between ranked choices. For instance, in the Olympic sport of archery, the point difference between first and second place may not be the same as the point difference between fifth- and sixth-place finishers. Their variability is not presumed by the ranking, but ranking occurs (i.e., first place is greater than second place).

Interval Measurement

If you add equal intervals between measurements to an ordinal scale, you have an interval scale. Interval measurement is defined as being measured by numbers in a more exact way than ordinal measurement. Specifically, interval data allow us to quantify differences between values (i.e., the difference between 1 and 2 is the same as the difference between 2 and 3). Interval data are particularly useful in social and behavioral sciences when employing questionnaires or surveys on attitudes, traits, or symptoms of, for instance, eating disorders, as our example study did. Ratings on a Likert scale are interval measurements as the differences between ratings are meaningful and the zero is arbitrary or meaningless (e.g., “Please rate each of the following statements on a scale of 1 to 5”). When interval measures include a zero, that does not mean that there is zero of that trait or characteristic (e.g., the interval measure of zero degrees Celsius does not mean there is no temperature at all). Unlike ordinal measurement, interval measurement can be added or subtracted in addition to being compared as greater or less than. In addition, interval scales may be constructed to allow for negative numbers – such as temperature. Ratio data allows averaging to calculate a mean, but because there is no true zero, ratio data does not allow multiplication or division.

Ratio Measurement

Ratio measurement is defined as being measured by numbers, similar to interval measurement, but with a true zero, which means there may be an absence of the measured variable. It includes weight, length, or width – all positive quantities with an absolute zero. Ratio measurement is not commonly used in social and behavioral research as you are unlikely to measure social/psychological constructs with a true absence of that characteristic. However, research using measures such as computer-key press reaction time or number of words spoken certainly does – true zeros are possible when the key is not pushed or no words are spoken.

While both interval and ratio scales result in data that can be added, subtracted, ordered, and summarized by a median or a mode, more can be done with ratio data. The true zero and equal intervals of ratio scales allow multiplication, division, and averaging scores to determine the mean. In addition, ratio scales allow comparisons, such as “three times as much” or “half as many” and also means that ratio measurement does not have negative values.

Frequencies

Frequency refers to how often something occurs over time. In the context of descriptive statistics, frequencies are numbers summarizing the data produced by a study. They can be displayed to show numbers or percentages of occurrences, and these can be displayed as histograms, bar graphs, line graphs, or pie charts. Sophisticated statistical programs are not required because a spreadsheet program (e.g., Excel) can be used to display the numbers, percentages, or proportions in tables, and the tables can be used to generate a variety of charts. Percentages can be calculated easily by hand, a spreadsheet, or a calculator.

Frequencies can summarize information about the participants or about the participants’ responses. For example, participants might be asked to indicate demographic information about themselves, such as whether they are female, male, or prefer not to identify their gender (nominal measure), their age (ratio measure), or their ethnicity (nominal measure). Frequencies that display participant responses communicate information about the variables. For example, a researcher might provide the number of participants who said they first tried alcohol at age 13 or younger and the number who said they first tried alcohol after age 13. A researcher studying influences on time-of-purchase donations might give the number of participants who donated when asked if they would like to round up the cost of their purchase and the number of participants who donated when asked if they would like to donate. A researcher studying the influence of how requests are worded on how people respond to those requests might study how people waiting in line to pay for items react to a stranger who asks to cut in line with a good reason, no reason, or a nonsense reason that is worded like a typical good reason (using “because” in the request).

Histogram

Histograms display frequency data by showing the frequency (i.e., number) of responses or observations for numerically ordered variables. The number of observations for each possibility is shown in a bar, and the bars touch one another. In a research study surveying 100 college women, all 18–23 years old, the number of people you interviewed at each age can be displayed in a histogram, as shown in Figure 22.1. The histogram shows the number of people at each age who completed the survey.

Shapes of Distributions

One of the most important functions of a histogram is that it allows you to describe the shape of a distribution for interval and ratio variables; this includes a statement about where data may cluster and how they spread out. An example of the most common shape in research is shown in Figure 22.1 – a *normal distribution* (i.e., the bell curve); this is a roughly symmetrical graph with most scores falling in the middle (i.e., the *body* or *peak*) and declining toward each side (i.e., the *tails*). However, not all interval or ratio data have that quality. The main alternative is a skewed graph in which most values are

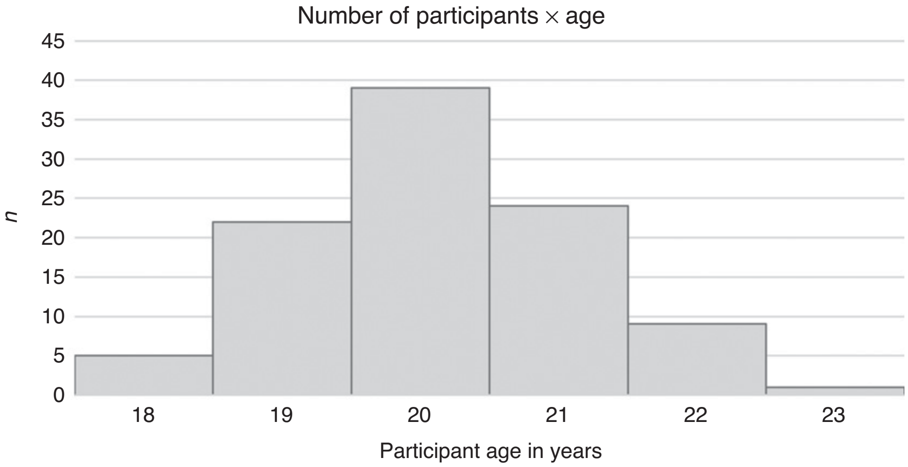


Figure 22.1 *Normal distribution for the number of people in each age group who completed the eating-disorders survey.*

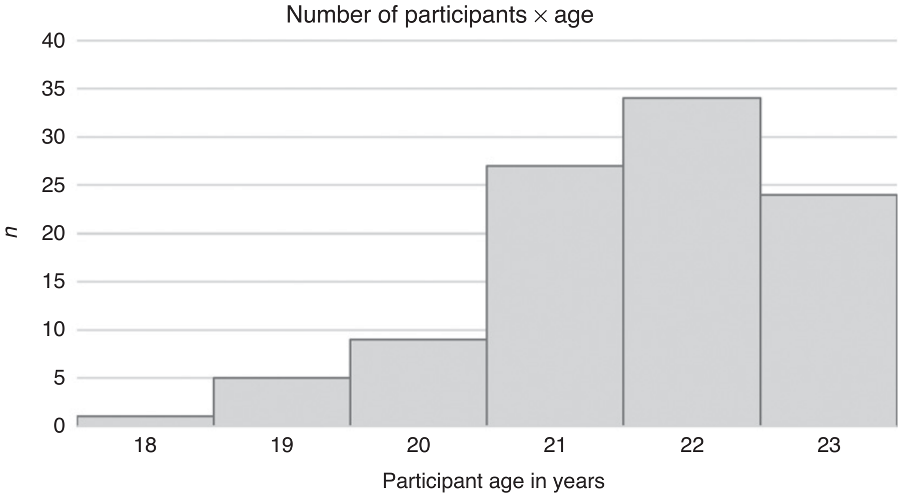


Figure 22.2 *Negatively skewed distribution for number of people in each age group who completed the eating-disorders survey.*

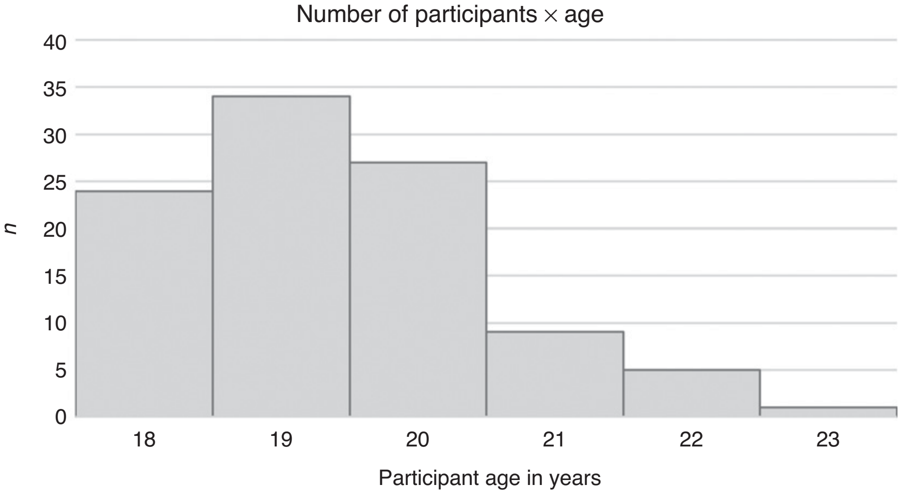


Figure 22.3 *Positively skewed distribution for number of people in each age group who completed the eating-disorders survey.*

on one side of the graph, with one of the tails being much longer than its opposite. Figure 22.2 illustrates a negatively skewed distribution, with the long tail on the left side of the curve, and Figure 22.3 shows a positively skewed distribution with the elongated tail to the right. The reason these shapes are so important is that they have an impact on which descriptive statistics methods we use, as discussed in the next section.

Bar Graphs

Bar graphs, like histograms, display the number of observations for each of the nominal measures (Smith & Davis, 2010). Each bar represents a category, and the bars could be re-ordered because the categories do not have a numerical order. Bar graphs show the frequency (or number) obtained in the study on one axis and the variables (the nominal categories) for which frequencies are shown on the other. The bars can be vertical or horizontal, though vertical is more typical. Unlike histograms, the bars do not touch.

Consider the demographic data mentioned earlier; the gender of participants can be shown in three bars with the number of participants who answered each question indicated along one axis, and the number of people who gave each answer along the other axis. However, there is no numerical order to the gender variable answers of male, female, or prefer not to answer. Imagine you tested 500 participants in a study of disordered eating, and 282 participants indicated they were female, 197 indicated they were male, and 21 indicated they preferred not to identify their gender. You could summarize this data visually by displaying the number of people who answered female, male, prefer not to say as shown in the two bar graphs in Figure 22.4. Alternatively, you could summarize the same data in a bar graph showing the percentage of 500 participants who answered female, male, or prefer not to say, as shown in Figure 22.5.

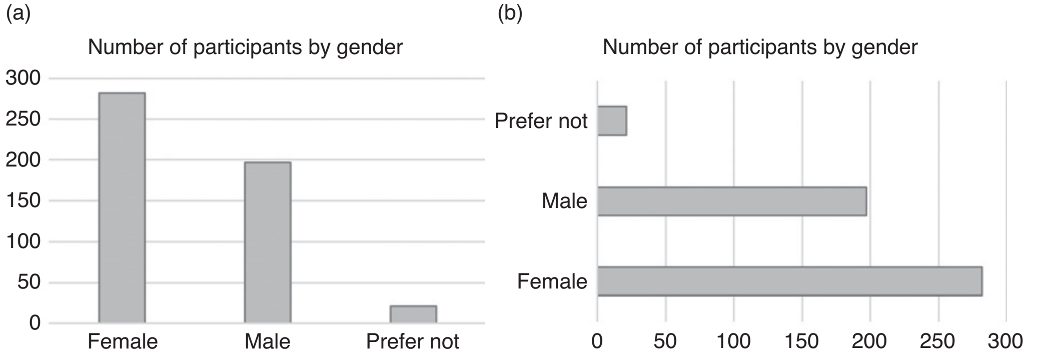


Figure 22.4 *Vertical and horizontal bar graphs of the number of participants by gender.*

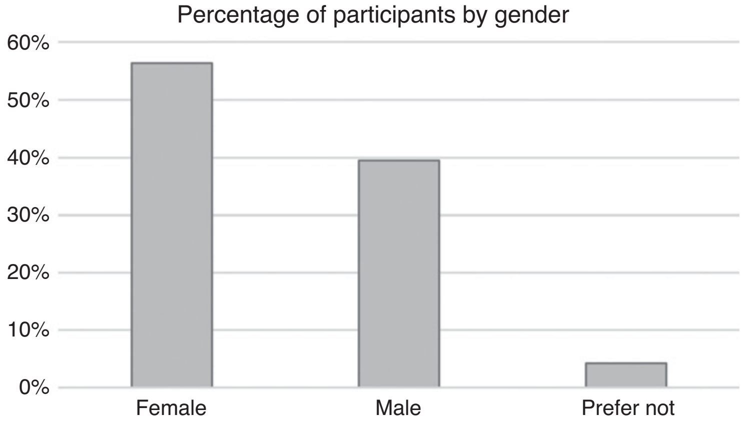


Figure 22.5 *Vertical bar graph of the percentage of participants by gender.*

Turning again to an example of an eating-disorders survey, both Figures 22.6 and 22.7 show the number of participants in each disordered-eating category. However, bar graphs may be clustered to show more information in additional bars. For example, Figure 22.7 shows the number of females, males, and those who preferred not to give their gender for each of the disordered-eating categories.

Because 282 was the largest number obtained in this example study, the frequency axis in Figure 22.4 only shows a maximum of 300, rounding up to the nearest hundred. However, be cautious about the highest frequency shown on that axis because increasing or reducing it could give the false impression that the frequencies are much more similar or different than they actually are. A longer axis (i.e., with higher numbers shown) may obscure frequency differences, while a shorter axis (i.e., with lowest number obtained shown) may exaggerate differences.

Consider the initial impressions of the two graphs in Figure 22.8 – displaying the exact same means and variables – for a hypothetical study measuring college student

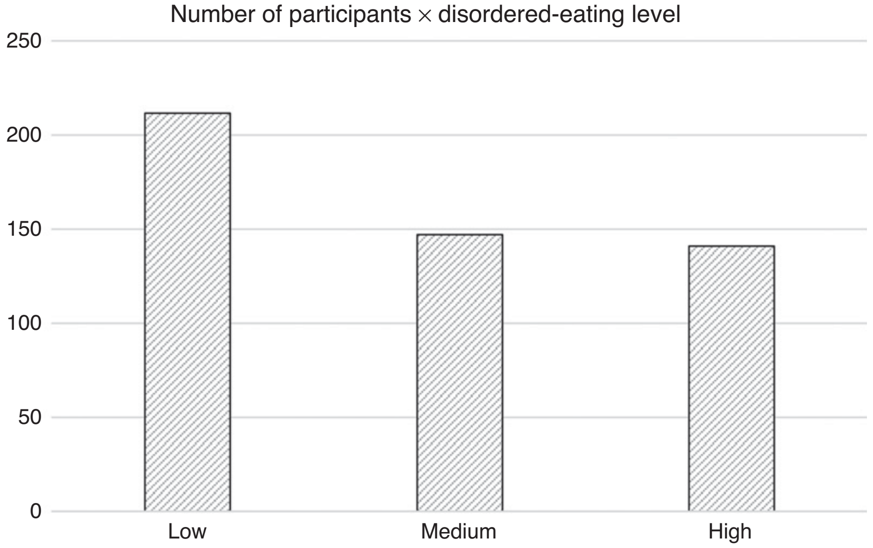


Figure 22.6 *Number of participants in low, medium, and high disordered-eating-level categories.*

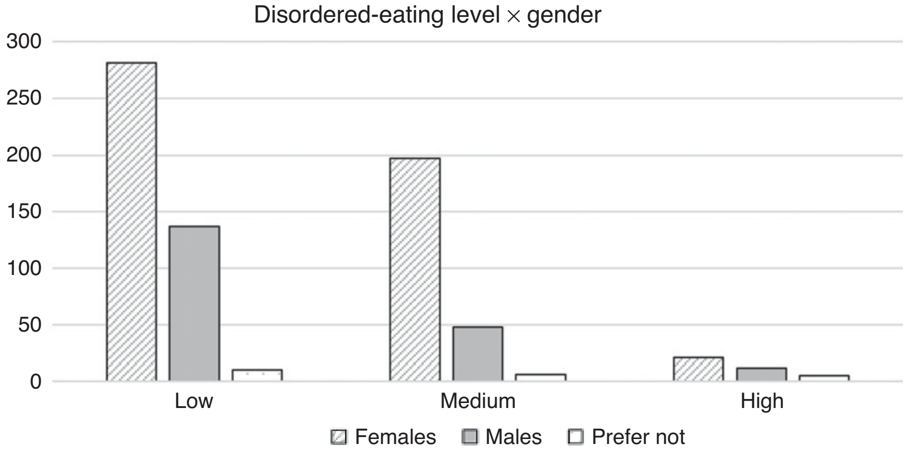


Figure 22.7 *Number of participants in low, medium, and high disordered-eating-level categories by gender.*

motivation on a 30-point scale and ethnicity. The first bar graph shows the maximum on the *y*-axis at 24, automatically generated by Excel based on the means. The second shows the *y*-axis changed to the survey maximum score possible of 30. The first bar graph gives the impression of larger mean motivation score differences between ethnic groups than the second bar graph. This example also demonstrates why inferential statistics are needed to answer the question "are there any *significant* differences between the groups in motivation scores?"

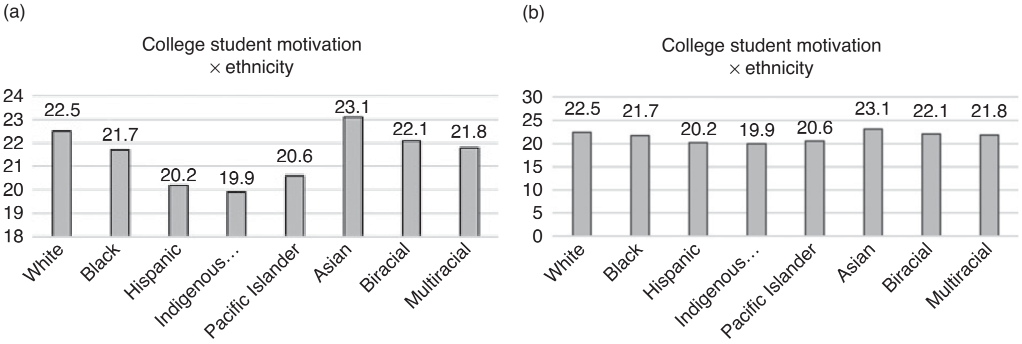


Figure 22.8 *Demonstration of the effect of a shorter or longer* y-*axis on bar-graph displays.*

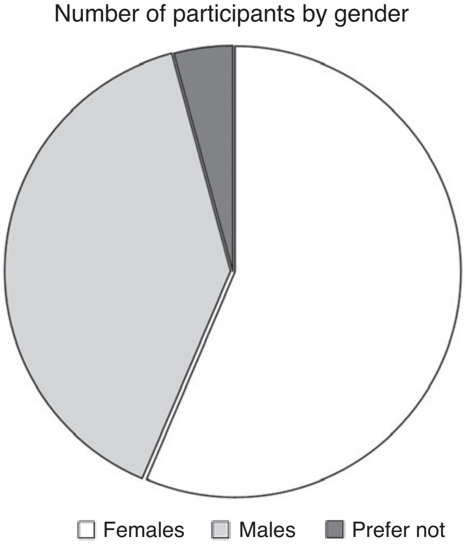


Figure 22.9 *Number of disordered-eating survey participants by gender.*

Pie Charts

Pie charts provide another method of displaying descriptive data. The pie chart in Figure 22.9 presents the same information on gender as the bar graphs in Figure 22.4.

Measures of Central Tendency

Measures of central tendency provide a numerical value that is representative of a center point of the sample. Three types of measures of central tendency are commonly used – mean, median, and mode; the best choice will depend on your collected data. Each is informative about what is common in your data. They can also be used to understand groupings within your data (e.g., center points for groups based on the country of origin). They also are often the basis of many other statistical tests (e.g., *t*-tests), which will be expanded upon later this chapter.

Mean

The mean is the average of data points within a variable and is one of the most commonly used measures of central tendency. However, it cannot be used to meaningfully calculate the central tendency of a nominal variable and is not generally recommended with ordinal variables. To calculate the mean, you add together all the values for the variable and divide by the total number of values. For example, if you have 20 participants reporting their pain on a scale of 1 to 10, the mean is calculated by summing each of their individual scores and dividing by 20 (the number of participant reports). Researchers often report means in their results sections, along with additional information about the variability of the scores. Table 22.1 provides an example of how the means in our example (Smith & Ransford, 1999) could have appeared in a table.

Median

The median is the middle value within a set of scores. The median is calculated by ordering all of the values in numerical order (i.e., from smallest to largest) and finding the middle value. If there is an even number of observations, it is the average of the middle two values. The median is useful in cases where there are outliers – a very small number of unusually high or low scores – in the data; outliers normally skew or push the mean in one direction more strongly than another, resulting in a measure of central tendency that isn't as representative of the data. The median is not always as easily used for comparison as the mean in statistics, but it can be used with ordinal, interval, and ratio data types. For example, the number of sexual partners can often be affected by outliers (e.g., individuals reporting high numbers of sexual partners) such that the median better represents the overall data than the mean.

Mode

The mode is the value that is found most frequently for a given variable. The mode is calculated from the frequency with which a value occurs within the data set and is

Table 22.1 *Conformity and eating-disorder inventory score (see Garner et al., 1983) means and standard deviations by sorority membership*

	Sorority		Non-sorority	
	Mean	SD	Mean	SD
Conformity	8.39	5.28	5.85	3.12
Eating-disorder inventory	50.5	40.93	31.5	23.37

particularly useful to describe nominal data. For example, if you asked participants what their primary language was, the mode could inform you which language was the most frequently reported in your sample. The mode is useful because it is not impacted by outliers, can be used with different scales of measurement, is the only measure of central tendency for nominal data, and can be used with more complex qualitative data. In some variables, there may be more than one mode or no mode at all (i.e., no two participants reported the same value).

Measures of Variability

Central tendency is a great way to summarize where the data are, but it does not tell the whole story. Imagine, for example, that you are reviewing examination scores from two classes. After doing a little research, you find that Class 1 has an average grade of 80%. Coincidentally, Class 2 also has an average of 80%. They appear to be equal, yet, upon close inspection, individual students in Class 1 vary quite a bit, with some earning 60% and others getting 100%. Class 2 students, on the other hand, tend to be more consistent, with individuals mostly earning 75 to 85%. For both classes, the mean is 80, but Class 1 has more variation.

This example illustrates the concept of *variability*, the degree to which individuals in a sample are dispersed nearer to or farther away from a measure of central tendency. It is common to see this on a histogram where the mean is used as the measure of central tendency; distributions with high levels of variability are spread out wide with many scores far from the mean. Low variability is noticeable because most of the individual scores are rather close to the mean. Of course, relying on just a visual assessment of variability does not give us much precision, so we must quantify it when dealing with interval- and ratio-level data.

Range

The range is the most basic measure of variability. It is simply the highest score minus the lowest score in the distribution. This is used mostly to illustrate how high and low the scores can go, but it fails to capture what is typical (i.e., how far from the center is the typical individual?). The interquartile range (IQR) is often used when the median is provided for central tendency, particularly when the distribution is skewed or there are outliers. As you recall, the median is the middle value within a distribution (i.e., the 50th percentile). Now, imagine you find the median on each side of the median; this would divide the distribution into four equal sections, with splits occurring at the 25th, 50th, and 75th percentiles. The IQR is found by simply subtracting the score at the 25th percentile from that at the 75th percentile. Taken together, the median and IQR tell us the median of scores and the median distance of an individual from the center (i.e., the median); the IQR tells us where the center 50% of all individual scores fall, centered around the median.

Standard Deviation and Variance

Two other measures are used far more often, however, because they go well with symmetrical distributions, and they relate to the mean as a measure of central tendency. These measures are the *standard deviation* – the average distance of an individual from the mean – and *variance* – the squared standard deviation (i.e., the average *squared* distance from the center). Table 22.1 shows how the standard deviation (SD) may appear in articles along with the mean.

Regardless of the statistic used, all measures of variability represent the same thing – the extent to which individuals in a distribution spread out around the measure of central tendency. Therefore, the higher the values in these statistics, the wider the histogram will be, the greater the differences among individuals, and the less consistency there is in measurement.

Correlations

Correlations display how two continuous variables co-occur – whether higher scores on one variable occur with higher or lower scores on the other one. Imagine collecting two scores for every participant and then plotting one point to represent each participant's two scores (i.e., for each participant, their scores on the *x*- and *y*-axes are represented by one dot). Once that has been done for all participants, the relationship between the two measures might reveal a correlation between the two variables you measured. For example, a researcher might measure how many times per week participants weigh themselves and how overweight they are; two numbers are recorded for each participant, then the relationship between the two measures is calculated for the entire group of participants. The strength of a correlation between the two variables is shown by a correlation coefficient (*r*), that ranges from −1.0 to +1.0. The closer to zero the correlation coefficient is, the weaker the correlation; the closer to −1.0 or +1.0, the stronger the correlation.

Correlations do not explain causal relationships, but they can spur researchers to develop experimental research projects that do address causation. This chapter will not be the only time you hear "correlation does not equal causation!" Remember, no matter how well you might be able to predict the score on one variable if you know the score on the other, you still cannot say anything about causation.

Positive Correlations

When the pattern of observed numbers reveals that as one variable increases, so does the other (and when one variable decreases so does the other), this is a positive correlation. A perfect positive correlation is represented by a correlation coefficient of 1.0. Consider the case of attractiveness and liking. If when participants rate others higher on attractiveness, they also rate them higher on likability, and when they rate others lower on attractiveness, they also rate them lower on likability, that relationship reflects a positive correlation. Based on this positive correlation only, you cannot say that being more attractive causes others to like more attractive people

https://doi.org/10.1017/9781009010054.033 Published online by Cambridge University Press

more and to like less attractive people less. The relationship *could* mean that is true, but it could also mean the opposite – people who are liked more are viewed as more attractive; it is possible that greater liking causes people to rate others as more attractive. It could also mean that some third, unmeasured variable causes or affects the relationship between attractiveness and liking. However, another possibility is that, even though the two variables are positively correlated, the variables may have no causal relationship whatsoever – it is a *spurious correlation*. This is why correlation does not equal causation; there could be a causal relationship, but you have no way of knowing which variable causes a change in the other(s) or if there is no causal relationship between the variables.

Negative Correlations

When the pattern of observed numbers on two variables reveals that as scores on one variable increase, scores on the other decrease, this is a negative correlation. A perfect negative correlation is represented by a correlation coefficient of −1.0. With a negative correlation, you know that if the score is high on one variable, it will be low on the other and vice versa.

No Correlation

Note that a negative correlation, which shows the two variables are related in a particular way, is very different from no correlation ($r = 0$); a negative correlation does not mean there is no relationship between the two variables. No correlation means that scores on the two variables do not occur together in any pattern. With no correlation, knowing the score on one variable does not give you any information about the corresponding score on the other variable.

Beginning the Transition to Inferential Statistics

Standardization and the *z*-Score

We have already seen that many variables take on a normal distribution when plotted in a histogram. The simplicity of this fact belies how incredibly useful that bit of knowledge can be. It turns out that, if you know the mean and standard deviation of a normal variable, you can do all sorts of things based on probability. At the most basic level, you can predict that 50% of all individuals will be above the mean and the rest below. But, as shown in Figure 22.10, you can extend this type of reasoning to standard deviations. Approximately 34% of all individuals will fall between the mean and one standard deviation above it; another 34% fall between the mean and one standard deviation below. This is true for *every* normally distributed variable, ranging from adult human intelligence quotient (IQ) scores to attitudes about eating disorders. In fact, if you are willing to do a little calculus, you can take any

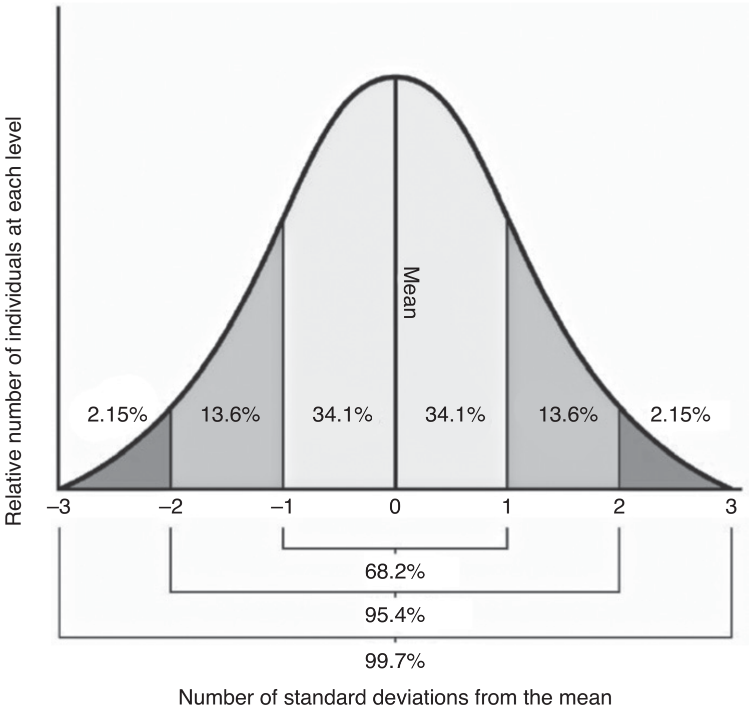


Figure 22.10 *Normal distribution and standard deviations.*

individual's score from a known population and, based on the mean and standard deviation, arrive at an estimate of their percentile rank – what proportion of scores is at or below that individual's own.

Because of the universality of the proportions found in standard deviations, statisticians often use a single variable, called a standardized score, when conducting statistical tests. The most basic of these is called a z-score – an individual's distance and direction from the mean as counted in standard deviations. The z-score does two useful things for us. First, it gives us a shortcut; once you know an individual z-score, you can skip the calculus mentioned above and look up the percentile rank associated with that value of z in a copy of the Unit Normal Table (*the z table*). You may also find it useful for making comparisons between different measures. For example, US high-school students generally complete either the ACT examination (mean = 21; SD = 5) or the SAT (mean = 1,000; SD = 200) to demonstrate college aptitude. If one student scores 26 on the ACT and the other scores 1,100 on the SAT, whose score indicates higher aptitude? In this case, the ACT score is $1z$ and the SAT score is $0.5z$, so the ACT score is better – it is further from the mean in the positive direction.

So far, our discussion has examined how individual scores, whether in populations or samples, can be described with distributions, central tendency, and variability. This is very useful information, but it gets even more sophisticated. Imagine you know the mean and standard deviation of your sample, but you want to make an educated guess about the mean of the population. You may also be working with a population that you understand well, but you want to make predictions about what a random sample might look like. Both of these tasks are possible, as long as you know something about how the process of sampling works. For this, we turn to the central limit theorem.

The *central limit theorem* (CLT) is a statistical theory about how samples work; given a few pieces of information about a population, you can make quite good predictions about the mean of the next sample you take. The foundation of the CLT is the concept of a sampling distribution. Whereas populations and samples are distributions of individual scores, sampling distributions are distributions of a statistic calculated from a specific sample size from a population. For example, imagine you have a known population of measures, such as IQ scores, with a mean = 100 and SD = 15. You could begin building a sampling distribution by taking a sample of $n = 25$ individuals, calculating the mean, and plotting it on a graph. After ensuring each sample of individuals has been returned to the population before selecting another random sample, you can do this again, and again, and again . . . ad infinitum. That is a sampling distribution. According to the CLT, we can know three things in advance about it:

(1) The mean of all sample means equals the population mean.
(2) The standard deviation of sample means is called the *standard error*, and it equals the population standard deviation divided by the square root of the samples' size.
(3) This sampling distribution is normal, especially when sample sizes are large ($n > 30$), and the original population is also normal.

In case you did not notice it, these three qualities are the ones we talked about in terms of descriptive statistics earlier; the only difference is that this describes how sample means are distributed (in terms of shape, central tendency, and variability) instead of individual members of a population or sample. That's important because, once you get the basics of a sampling distribution, you can then employ *z*-scores to make predictions about finding the likelihood of a specific sample mean taken from a known population. Alternatively, you can make *inferences* about a population from a single sample. This, in fact, is a whole new application of the CLT called inferential statistics, using sample statistics to make inferences about the value of a population parameter.

Confidence Intervals

The CLT is powerful because it allows us to make predictions about statistics (e.g., means calculated from a sample) and inferences about parameters (e.g., means calculated from the entire population, rather than just a sample). The main tool for both of these tasks is called a confidence interval, a combination of a point estimate and a symmetrical margin of error used to estimate a population or sample mean. To illustrate this, imagine you are sampling from a known population – adult IQ scores with a mean of 100 and SD of 15 points. We want to know where our sample mean will be if we take a random sample of 25 individuals from the population. We will use the following steps to create a confidence interval for our prediction:

Make a point estimate. The point estimate draws from the CLT; we know that the mean of sample means is equal to the population mean. Therefore, we should expect our sample mean to equal 100 in this case (the population mean).

Determine the standard error. We know that the sample mean will not be exactly equal to the population mean thanks to random error. The CLT theorem tells us to expect the mean to vary by the standard error (SE) on average – in this case $\mathrm{SE} = 15/\sqrt{25} = 3$.

Establish a confidence level. Confidence levels are based on proportions or percentages and are usually very close to 1.0. The default confidence level is 95%, but that is simply based on custom. To be more confident, you can go to 99%. However, there are drawbacks to this as you will read later in the section on hypothesis testing. For computations, the confidence level is expressed in z-scores, and the z-score for a 95% confidence level is ±1.96.

Build the confidence interval. With all of this information at hand, we are ready to build the confidence interval. We multiply the confidence level by the SE and add that to the point estimate. It is important to remember that the confidence interval is symmetrical, and it comes in both a positive and negative value (±1.96). This produces a lower and an upper boundary to our confidence interval. In this case, those values are 94.12 and 105.88.

Now that we have our confidence interval, we can phrase our prediction – we are 95% confident that our sample mean will fall between 94.12 and 105.88. When we take our next sample, will we be correct? Usually we will be, except for the 5% of the time when we get an unusual sample that falls outside of the confidence interval (see Figure 22.11).

This example works from a known population to an unknown sample. The same basic procedure is used much more often in the reverse direction – you have a sample that you gathered in research and need to make an inference about the population it

95% Interval

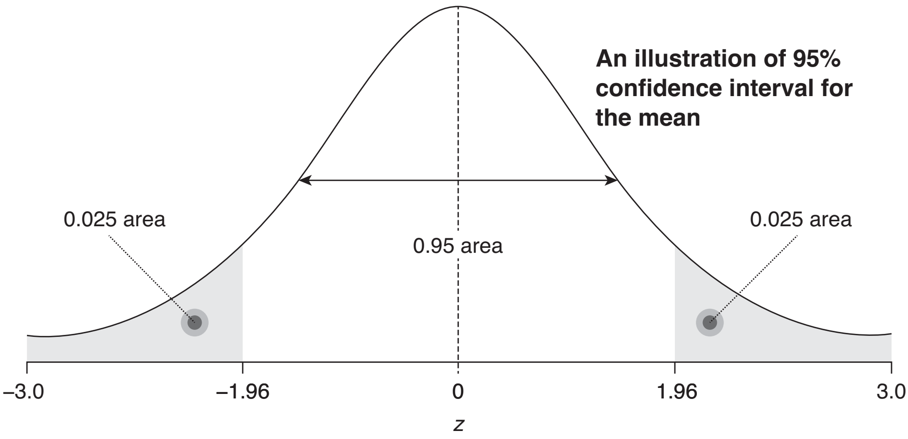


Figure 22.11 *An illustration of 95% confidence interval for the mean.*

comes from. For example, imagine a psychologist has developed a personality rating scale to measure a trait they call cross-cultural openness. Based on a sample of 225 students from the university, we know that the sample mean is 20 with a standard deviation of 5. However, we do not know what the population mean is, but a confidence interval can help. Working in the reverse direction, we can use our sample mean as the point estimate and calculate $\mathrm{SE} = 5/\sqrt{225} = 0.33$. Calculating a confidence interval (20 ± 1.96 × 0.33) suggests we can be 95% confident that the population of student scores at this university is between 19.3532 and 20.6468.

Hypothesis Testing

The purpose of statistical testing is to tell us whether our inference or comparison is more likely than no inference at all. The hypothesis researchers hope to support is the alternative hypothesis – that there is a relationship or a change, effect, or difference between groups. The null hypothesis is that there will be no relationship or change, effect, or difference between groups. Once the statistical test is conducted, taking the appropriate alpha level into account, the researcher decides whether or not to reject the null hypothesis.

In hypothesis testing, just as in any testing, it is important to understand the potential rate of false results. Errors in hypothesis testing are the likelihood that the outcome of statistical testing does not match the correct inference.

Type I Error

A *Type I error* (see Table 22.2) occurs when the null hypothesis is rejected when it should not have been rejected. For example, if a researcher runs a *t*-test to compare a group that received a treatment to a group that did not receive treatment, a Type I error occurs if the statistical significance indicates that the treatment was not more effective than the non-treatment group, when in fact it was. It is impossible to completely remove the possibility of Type I error, but it can be greatly limited by the significance level required to decide that the result is statistically significant. Following a general rule of thumb, researchers have accepted that 5% probability (*p*-value) is appropriate for limiting Type I error, known as the statistical significance level (alpha). The *p*-value of a statistical test is the probability of observing the same result or something more extreme, given the null hypothesis is true; thus, lower *p*-values indicate a lower probability that a Type I error will occur. Other researchers prefer to reduce Type I error further by determining the statistical significance when the *p*-value is less than 1%, but there is no way to completely eliminate the possibility of falsely rejecting a null hypothesis.

Table 22.2 *Chart of Type I and II errors*

	Null hypothesis is true	Null hypothesis is false
Reject null hypothesis	Type I error	Mostly correct conclusion
Fail to reject null hypothesis	Mostly correct conclusion	Type II error

Type II Error

A *Type II error* (see Table 22.2) occurs when the null hypothesis is not rejected when it should have been rejected. In the same example above, comparing a treatment group to a non-treatment group, a Type II error would occur if the researcher found no statistical difference when there was a true difference in group outcomes (see Figure 22.12). One way to reduce Type II errors is to increase the alpha level required to reject the null hypothesis; however, as you likely recognize, that increases the likelihood of a Type I error. However, there are other ways to reduce Type II errors.

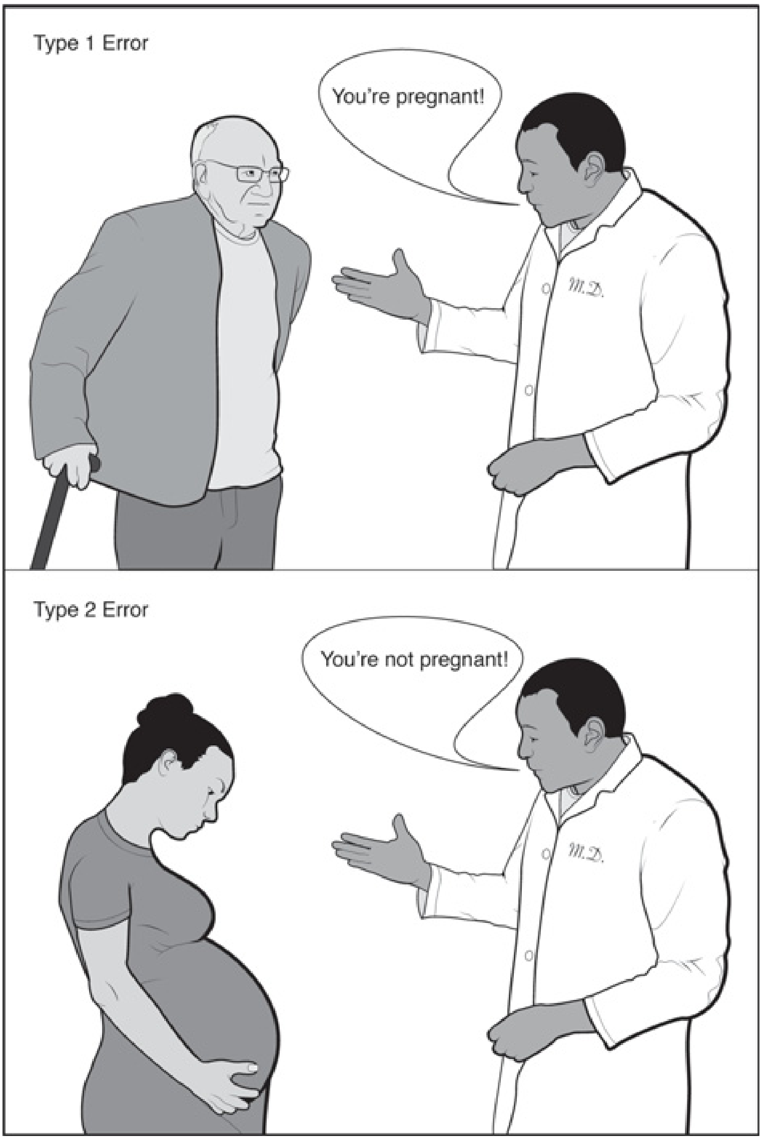


Figure 22.12 *Examples of Type I and Type II errors.*

Just as a Type I error is partially determined by the alpha level, a Type II error is determined by the power (*β*) – the likelihood of detecting a true alternative hypothesis. The power is determined by several factors outside of the alpha level. Specifically, the power is also impacted by the size of the sample in the data set and the effect size. Effect sizes are quantified by the size of the inference (e.g., the size of the difference between two groups) and can be used to determine the power a researcher has in testing a particular hypothesis. Specifically, larger effect sizes increase the power in hypothesis testing. Similarly, the larger the size of the sample, the more power there is to detect even small effect sizes without reducing alpha levels.

Hypothesis Testing: The z-Test

With knowledge of confidence intervals and a little bit of basic logic, scientists can conduct statistical tests that are able to detect changes or differences between a sample of interest and the status quo – what is normally expected of a sample. Collectively, these tests are known as null hypothesis significance tests (NHSTs), a set of techniques that are used to infer if differences between samples are likely to be the result of real, scientific phenomena or just random variations within samples. We can illustrate the concept of a NHST with the *z*-test, a way of determining whether a single sample is representative of a known population. However, as you will read, there are many other situations to which the basic logic and procedures of NHSTs can be applied.

Any NHST can be broken down into five steps. Let's imagine the known heart rate for adults in their 20s is 69 beats per minute with a standard deviation of 12. Let's test to see whether an intervention – having 16 college students climb two flights of stairs – will change that heart rate. Here's how we would analyze that study using a form of NHST called a *z*-test.

Step I. Establish a null hypothesis. The null hypothesis is a way of saying what is known about a population. Further, if we take a sample from that population, nothing (i.e., "null") should make that sample have an unusual mean. In our example, the null hypothesis is that, according to the CLT, our 16 students should have an average heart rate of 69, the same as the population mean. However, as researchers, we are often trying to gather evidence that something has changed; we will create an alternative hypothesis – the sample of students who climb two flights of stairs will not have a mean heart rate of 69. Notice that the null expresses what is known, and it is the second, alternative hypothesis that is hypothetical – *if* people climb stairs, *then* their heart rate will change.

Step II. Set criteria for differences. The CLT (specifically, standard error) tells us that, even if our sample of students does nothing, their mean heart rate is unlikely to be exactly the same as the population – 69 beats per minute. But how far from the population mean does your sample have to be before you are willing to say it is different? 71 beats? 75? To formalize this process, we can establish a criterion based on probability. It usually involves building a 95%

confidence interval around the population mean, leaving you a 5% chance of error in your decision-making process. As mentioned in the previous section, a 95% interval corresponds to a z-score of ±1.96. To say there has been a real, meaningful change in mean heart rate, our sample mean has to fall outside of our 95% confidence interval for the population mean.

Step III. Calculate the test statistic. The first two steps are based in logic. Step three is arithmetic, and it involves calculating the mean of the sample and converting it to a z-score. Let's say that we obtained a sample mean of 84 from our students who walked up two flights of stairs. Using the z-score formula, we find that the sample mean has a z-score of 5.

Step IV. Make a decision. If the sample mean lands within the confidence interval, we can say that it is different from the population mean but no more different than we would expect by random chance. However, if it falls outside of the confidence interval, it meets our criteria for *statistical significance*, being further from the mean than one would expect by chance. What happened in our example? Our 95% confidence interval ranged from $-1.96z$ to $+ 1.96z$, and our sample mean fell outside of those limits at $+5z$. We can say that the null hypothesis is false and accept the alternative.

Step V. Interpret. First, we need to determine what it means to falsify the null hypothesis. Because our sample mean was so high (measured in z-scores) it is very unlikely that it occurred by chance. In other words, it is unlikely that just by some weird, random chance we selected a number of students with tachycardia, a clinically elevated heart rate. It is probably the experimental manipulation (climbing stairs) that caused the difference. However, this only tells us that the difference is probably real, not whether it is important. Therefore, we need to have some idea of the impact of the manipulation – the *effect size*. The measurement of effect size changes according to different versions of NHSTs but, for z-tests, one of the most commonly used is Cohen's d; it measures how many standard deviations are between the population mean and sample mean. In our case, Cohen's d equals 1 because the means were one standard deviation apart. Cohen's d is generally evaluated according to absolute values, as listed in Table 22.3; for our example, we have a large effect – climbing stairs has a big impact on heart rate.

Table 22.3 *Interpreting the effect size of a* z- *or* t-*test with Cohen's* d*

Absolute value of d	Size	Example
0.0–0.2	Small	Climbing half a flight of stairs might raise your pulse by 2–3 beats per minute
0.2–0.8	Medium	Climbing one flight of stairs might raise your heart rate by 5–9 beats per minute
0.8 and above	Large	Climbing two flights of stairs greatly raises your heart rate

*Adapted from Cohen (1992).

Inferential Statistics

Chi-Square Tests

What is quantitative psychologist Amanda Montoya's favorite quantitative method – and why? She answered:

> I was recently in a meeting, and we were all talking about how much we love χ^2 (chi-square) tests. They were the one test in intro stat that you actually felt okay calculating by hand. I'm also particularly attached to χ^2 tables because you can calculate almost any other statistical table from a χ^2 table, which I think is just so cool (Montoya & Cannon, 2019, p. 25).

So how do you go about calculating a chi-square test by hand? First of all, determine when it is appropriate to use it. Chi-square tests are used with categorical variables. They can be used to understand the distribution of one variable or to test the relationship between two variables. There are three types of chi-square tests. A *chi-square goodness of fit* determines if measures on one categorical variable differ significantly from expected scores. A *chi-square test for independence* is used to see if two variables occur independently of one another. Finally, a *chi-square test of homogeneity* is used to see if two samples are likely to have come from the same population. Chi-square tests are used to determine if the observed numbers obtained differ sufficiently to reject the null hypothesis – if the numbers differ by enough to say that there is an effect of the variable that was greater than chance and greater than small meaningless variations.

To test the null hypothesis, chi-square tests make use of expected values compared to measured values on the variables of interest. A simple table, called a contingency table, can be used to lay out the values the researcher observed and the expected values if the variable in question had no effect. The contingency table then allows you to calculate the chi-square, a number you need to decide if you can reject the null hypothesis. You make that decision by comparing your chi-square number to a chi-square table – just as you used a table to make your final decision when conducting a *z*-test.

The formula for chi-square is:

$$\chi^2 = \Sigma \frac{\left(O^2 - E^2\right)}{E}$$

It is actually a simple calculation – as Dr. Montoya intimated – especially if you use a spreadsheet to lay out your contingency table. For every observed value (*O*), subtract every expected value (*E*), square the results, then divide that total by the expected value, and add those values together. That's your chi-square value!

A concrete example of a chi-square goodness-of-fit test will demonstrate even better how easy it is to use this formula and introduce the basic concepts behind the chi-square test. Imagine we wanted to see if the number of men and women psychology majors at a university are the same proportion as men and women attending the university as a whole. If the university provides the information about the number of students on campus by gender, you have the actual counts for the population. Now you need to determine how many men and women major in psychology and how many you would expect if psychology majors reflect the same gender proportions as appear across campus. First, construct your contingency table, shown below as created in Excel (see Table 22.4). Students are given four options when asked to indicate their gender – male, female, non-binary, prefer not to answer – and the university provides the actual numbers. In addition, you know your *observed* numbers for each gender option – the 832 psychology majors' answers. What you need to do next is figure out what the *expected* number of psychology majors should be for each gender category.

To figure out the expected numbers for psychology majors, obtain the percentage of the university total of 12,391 students for each category (e.g., 5,500/12,391 = 0.44, so 44% of the 12,391 students are male). Then, determine the same percentage for each corresponding category based on the total of 832 psychology majors, as shown below (e.g., 44% of 832 psychology majors should be male, or 832 × 0.44 = 369 males); see Table 22.5.

To compare your score to the chi-square table, you need to know degrees of freedom (df) and to choose your *p*-value. The most commonly used *p*-values are 0.05 and 0.01, and df is determined by the number of categories minus 1 in your contingency table. In this case, four categories of gender – 1 = 3, so df = 3.

Table 22.4 *Contingency table example for initial chi-square calculations*

Categorical variable – with four categories	*O* (observed measures – psychology majors' gender)	University-wide student gender numbers	*E* (expected measures if null hypothesis is true)
Male	134	5,500	?
Female	655	6,723	?
Non-binary	32	111	?
Prefer not to answer	11	57	?
TOTAL	**832**	**12,391**	

Table 22.5 *Contingency table example for calculating expected measures and chi-square*

Categorical variable with four categories	*O* (observed measures – psychology majors' gender)	University-wide student gender numbers	*E* (Expected psych major numbers if null hypothesis is true)	$O - E$	$(O - E)^2$	Divide $(O - E)^2$ by E
Male	134.00	5,500.00	369.30	−235.30	55,366.23	149.92
Female	655.00	6,723.00	451.42	203.58	41,445.12	91.81
Non-binary	32.00	111.00	7.45	24.55	602.55	80.84
Prefer not to answer	11.00	57.00	3.83	7.17	51.45	13.44
TOTAL	**832.00**	**12,391.00**	**832.00**			**336.02**

According to the chi-square table, the critical chi-square level is 7.81; because 336.02 is greater than 7.81, we can reject the null hypothesis and conclude that the gender distribution of psychology majors is not like the gender distribution of the university.

t-Tests

The *z*-test is only one form of NHST. The same basic steps can be applied to various situations in which data might be collected differently. For example, if you have a known population but do *not* know the population standard deviation (we knew both in the *z*-test example above) you can solve the problem with a one-sample *t*-test. For this test, you would follow all the same steps except for two things. First, you would use the standard deviation as estimated from your sample and, second, because of this, you would use the *t*-statistic instead of *z*. The difference between *z* and *t* is simple; *z* represents a known value, but because *t* is an estimated standard deviation, it is adjusted to be larger than *z*. This allows for *t* to be a little more conservative and less likely to produce errors than if *z* were used with an estimate.

We also can use *t*-tests in situations where there are two samples. For example, an *independent samples t-test* is used to analyze true experiments with an experimental and control group. For our simple example, imagine a group climbing two flights of stairs while another group takes the elevator; this would constitute two independent samples, each with its own mean and standard deviation. In terms of an NHST, the same four steps apply with a few modifications. The null hypothesis is that the mean difference between the groups is zero (i.e., the two groups are the same), and the alternative hypothesis is that the mean difference between the groups is not zero. Although the second and third steps substitute *t* for *z*, there are only minor differences in how the calculations are done; the logic of the steps remain the same. In fact, the last two steps – decision-making and interpretation – are identical to the *z*-test.

Another two-sample *t*-test is known as either the *paired-samples t-test* or *repeated-measures t-test* (both names refer to the same statistical procedure). This analyzes the differences between group means when the individuals in each sample are somehow connected to each other. In a paired-samples *t* test, we might identify a student in the first group as a 20-year-old male with a resting heart rate of 70. To create the second group, we would just identify a second 20-year-old male with the same resting heart rate. By building demographically similar groups, we can rule out the influence of things like age and baseline heart rate as contributing factors when we calculate the *t*-test. Similarly, we can control for those factors by doing a repeated-measures test; we would have several students complete the experimental condition first, rest for 10 minutes, and then complete the control condition. This would allow us to compare one individual's scores from both conditions.

Again, the *t*-test is just a variation of the five-step NHST process, so we do not need to run through all of the steps again. However, it may be helpful to review a research article in which *t*-tests are used, such as the Wright et al.'s (2016) study of health behaviors. As you read the paper, you will see that it begins with a background of past research and a justification for the current project followed by a methods section that describes how data were collected. The results section provides descriptive statistics of the sample (means and standard deviations) and identifies the type of *t*-test used. In Table 1 of Wright et al.'s paper, you can find confidence intervals based on the difference between two means, *t*-test values, and whether the test is statistically significant. More importantly, you can read the text of the results section to see the authors address steps four and five of the NHSTs – the decisions and interpretations, including Cohen's *d*.

One-Way ANOVA

A one-way analysis of variance (ANOVA) is a statistical test that allows for the comparison of more than two groups. Whereas *t*-tests allow for a comparison between two groups on one independent variable, one-way ANOVA allows for a comparison among more than two groups on one independent variable. For example, a researcher wanting to compare means of altruism among gender groups (e.g., cisgender female, cisgender male, and non-binary) may want to use a one-way ANOVA. Groupings in an ANOVA must be categorical such as a nominal variable but can be created using ordinal, interval, or ratio data with more steps. The one-way ANOVA compares means of the groups on a particular dependent variable.

Independent Samples

An independent-samples one-way ANOVA has an alternative hypothesis that there are significant differences between two or more groups on the dependent variable. The null hypothesis states that there are no differences between groups. The *F*-statistic is used to determine the *p*-value (probability) of rejecting the null hypothesis. The *F*-statistic does not inform you about specific differences between groups but just that there is a difference between at least two of the groups in the independent variable. To determine specific differences, you must conduct post hoc or multiple comparison testing. Post hoc tests pair groups to compare – Group 1 is compared to Group 2, Group 1 is compared to Group 3, and Group 2 is compared to Group 3-in pairs. ANOVA post hoc testing is beneficial in that the overall error rate is controlled for in a way it is not in *t*-test comparisons. However, this also results in increased Type II error rates because more comparisons are conducted in post hoc testing.

The three most commonly used post hoc tests include Tukey's test, Holm's method, and Dunnett's correction. Tukey and Holm's tests compare every pair to each other, and Holm's test is slightly more protective against Type II errors than Tukey's test. Dunnett's correction is not used for comparing each group to the others in pairs but to compare each group to one control group. In the article on eating disorders among women in college (Smith & Ransford, 1999), they conducted *t*-tests because they had two groups (e.g., sorority and non-sorority women); however, if they added a third comparison group, they could conduct an independent-samples

ANOVA. Specifically, they might add women of a similar age range not currently attending college (i.e., college and sorority attending, college and non-sorority attending, and those not attending college or a sorority).

Correlated Samples

A repeated-measures one-way ANOVA tests more than two groups but assumes the individuals in each group are linked across groups in some way. This is most typically used when the same participants are included in each group. This is useful in examining differences across time. For example, people may participate in a training with more than three parts, where the creators of the training want to evaluate the learning from that training across time. The researchers evaluate the same participants across multiple time points – before the training, after each individual training, and at the end. Thus, they can use a correlated-samples ANOVA to evaluate if there is a significant difference in mean learning across multiple training time points. Another method is by utilizing a randomized block design – grouping individuals by a similar characteristic and placing them into different groupings. For example, perhaps a person in the counseling center decides to run interventions to reduce disordered eating in female college students. They select three methods of intervention: a do-it-yourself workshop online, an in-person workshop, and a worksheet they give to women on campus. They then randomly assign sorority and non-sorority members to the groups to determine if they impact the scores of the women differently. By collecting scores before and 30 days after completing the intervention they are assigned to, they can compare their scores and see if there was an impact by type of intervention.

Conclusion

This brings us to the end of this chapter's statistics journey – from descriptive statistics, through transitioning toward making causal inferences, to considering three basic inferential statistical methods. However, this chapter also can be seen as a beginning – the beginning of a journey toward understanding the many other methods of statistical analysis and how they can be used to reveal social science findings.

References

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. http://dx.doi.org/10.1037/0033-2909.112.1.155

Garner, D. M., Olmstead, M. P., & Polivy, J. (1983). Development and validation of a multidimensional eating disorder inventory for anorexia nervosa and bulimia. *International Journal of Eating Disorders*, *2*, 15–34.

Montoya, A. & Cannon, B. (2019). Practicing quantitative psychology (from an aerial circus trapeze!?) with Amanda Montoya, PhD. *Eye on Psi Chi*, *23*(3), 22–25. https://doi.org/10.24839/2164-9812.Eye23.3.22

Smith, N. N. & Ransford, C. (1999). The relationship between eating disorders and conformity in female college students. *Psi Chi Journal of Undergraduate Research*, *4*, *9–11*. Available at: www.psichi.org/resource/resmgr/journal_1999/Spring99_Smith.pdf

Smith, R. A. & Davis, S. F. (2010). Using statistics to answer questions. In *The Psychologist as Detective: An Introduction to Conducting Research in Psychology*, 5th ed. (pp. 171–202). Prentice Hall.

Wright. R. R., Broadbent, C., Graves, A., & Gibson, J. (2016). Health behavior change promotion among Latter-Day Saint college students. *Psi Chi Journal of Psychological Research*, *21*, 200–215. https://doi.org/10.24839/2164-8204.JN21.3.200

23 Testing Theories with Bayes Factors

Zoltan Dienes

Abstract

Bayes factors – evidence for one model versus another – are a useful tool in the social and behavioral sciences, partly because they can provide evidence for no effect relative to the sort of effect expected. By contrast, a non-significant result does not provide evidence for the null hypothesis tested. If non-significance does not in itself count against any theory predicting an effect, how could a theory fail a test? Bayes factors provide a measure of evidence from first principles. A severe test is one that is likely to obtain evidence against a theory if it were false – to obtain an extreme Bayes factor against the theory. Bayes factors show why cherry picking degrades evidence, how to deal with multiple testing, and how optional stopping is consistent with severe testing. Further, informed Bayes factors can be used to link theory tightly to how that theory is tested, so that the measured evidence does relate to the theory.

Keywords: Bayes Factor, Severe Test, Evidence, Multiple Testing, Optional Stopping, Cherry Picking, Priors

Introduction

An integral part of science is testing theories, and many journal articles are phrased as attempts to test theories (contrast with McPhetres et al., 2021). A key inferential tool used for this purpose is statistics – often null hypothesis significance testing (NHST). However, a non-significant result using the null hypothesis (H_0) of no effect does not in itself provide evidence against a theory that predicted an effect. That is, a theory that predicts an effect is not necessarily disconfirmed just because the outcome was non-significant (Edlund et al., 2021); paradoxically, it may even receive confirmation (Dienes & McLatchie, 2018). So, how can we put our theories to the test – ideally, to a "severe" test – where there is a chance they may fail if they are false? After all, passing a severe test would intuitively seem to provide evidence for a claim. Consistent with this, a Bayesian approach is to first define evidence; the strength of evidence of data for one hypothesis versus another is the amount by which the strength of belief in the one hypothesis versus the other should change in the light of the data. This definition, combined with the claim that the strength of belief should ideally be consistent with the axioms of probability, leads to the Bayes factor. This chapter will go through how the Bayes factor as a measure of strength of evidence can be used to severely test theories.

Bayes Factors as Evidence

How much should strength of beliefs change in the light of data? Let $P(\mathrm{H}_1)$ be the probability of H_1 – the strength of belief in H_1 – where H_1 is e.g. a hypothesis that something exists (e.g., a difference between conditions and a relationship between two variables). Let $P(\mathrm{H}_0)$ be the probability of H_0 – the strength of belief in H_0 – where H_0 is the hypothesis that e.g. something does not exist (there is not a difference between conditions; there is no relationship, etc.). For example, H_1 could be the hypothesis that there is subliminal perception in a certain paradigm; H_0 that it does not exist. $P(\mathrm{H}_1)/P(\mathrm{H}_0)$ is called the prior odds in favor of H_1 rather than H_0 – the relative strength of belief in H_1 rather than H_0 before data are collected. Data D are collected. Then $P(\mathrm{H}_1|\mathrm{D})/P(\mathrm{H}_0|\mathrm{D})$ is the posterior odds in favor of H_1 rather than H_0 – the relative strength in belief in H_1 rather than H_0 in the light of data D. With some simple rearranging of the axioms of probability, one obtains

$$\begin{aligned}&P(\mathrm{H}_1|\mathrm{D})/P(\mathrm{H}_0|\mathrm{D}) = P(\mathrm{D}|\mathrm{H}_1)/P(\mathrm{D}|\mathrm{H}_0) \times P(\mathrm{H}_1)/P(\mathrm{H}_0)\\ &\text{Posterior odds} = \text{Bayes factor} \times \text{prior odds}\end{aligned}$$

Thus, the Bayes factor is the amount by which one should normatively change confidence in one hypothesis versus another. That is, the Bayes factor is a measure of the strength of evidence from first principles – the axioms or probability and a definition of evidence (Jeffreys, 1939; Morey et al., 2016; Rouder et al., 2009).

Prior odds may be purely personal. I may believe a priori that subliminal perception is quite likely. Someone else may believe that subliminal perception is scarcely credible. Yet, if we both agree on the predictions made by the theory that perception is subliminal under stated conditions, and if these predictions can be made for objective reasons (as we discuss later), then the $P(\mathrm{D}|\mathrm{H}_1)$ and $P(\mathrm{D}|\mathrm{H}_0)$ are relatively objective. The Bayes factor tells each of us how much to change our beliefs in the light of data, in the same direction for both of us, even if we began at different starting points. We started this section by interpreting probabilities as subjective (i.e., as strengths of belief). By doing so, we can separate out what may be purely personal (i.e., prior odds) from the relatively objective message of the data – the Bayes factor.

If the data were impossible on H_1 and possible on H_0, the Bayes factor would be zero – the data would provide zero evidence for H_1 rather than H_0 (or conversely, infinite evidence for H_0 rather than H_1). H_1 would be falsified. Conversely, if the data were impossible on H_0 and possible on H_1, the Bayes factor would be infinite, and the evidence would be infinite for H_1 relative to H_0. H_0 would be falsified. If the data were equally probable on H_1 versus H_0, the Bayes factor would be 1, and the data would not be evidence for either hypothesis relative to the other. In sum, the Bayes factor varies between 0 and infinity, with 1 as the neutral point of no evidence.

How far from 1 does a Bayes factor have to go before the evidence is good enough for a decision? If there are clear costs and benefits, the answer can be rationally determined (Lindley, 2014). However, in many cases in the social and behavioral sciences, where a decision needs to be made, determining the costs and benefits is a difficult and unilluminating extra step. Has a possible confound been ruled out clearly enough that we can move on and apply the main theory to another problem? Has a version of the main theory been corroborated well enough that for now we continue to work on it rather than postulate another variant? Should a variable be dropped from the model? In most cases, and in most areas of the social and behavioral sciences, the prior probabilities and costs and benefits are sufficiently similar for the options considered that a convention for good enough evidence would be useful. A conventional degree of good enough evidence would also help stop special pleading for different strengths of evidence for different decisions based on how well it suited a researcher. However, a Bayes factor is a continuous measure of the strength of evidence without special bumps at particular values, and it is always open to a researcher to argue against a convention in particular cases.

In terms of the history of statistical inference, Edgeworth in 1885 suggested that two standard errors' difference was evidence for a difference between means "just worth taking note of" (Stigler, 1999, p. 103). According to Cowles and Davis (1982), in the early years of the twentieth century, two standard errors difference had become a frequent convention for significance in a variety of sciences. Thus, Fisher (1925) followed an already existent tradition in explicitly proposing $p = 0.05$ as a suitable significance level – the p-value that roughly corresponds to two standard errors difference. Jeffreys (1939) developed Bayes factors as a system of inference. He found that, when he addressed the same problems as Fisher, a Bayes factor of three roughly corresponded to 0.05 (i.e., 5%) significance. This is true, provided the obtained difference (or sample parameter value more generally) is roughly that expected based on H_1. Thus, using a Bayes factor of roughly 3 as a convention for evidence "just worth taking note of" is not arbitrary; it is roughly the amount of evidence scientists have been using for 100 years. That does not mean it is useful to take the convention as a black and white boundary. Stigler's phrase "just worth taking note of" is often the right attitude. And one may wish to revisit conventions no matter how ancient they are. Cortex (2021), for example, now uses a convention of 6 for a Bayes factor indicating good enough evidence (and correspondingly a significance level of 2%).

If a Bayes factor simply re-expressed the evidence already represented by a given significance level, not much would be gained by using Bayes factors. However, there is no monotonic relationship between a Bayes factor and a p-value (Jeffreys, 1939; Lindley, 1957, Morey, 2018). The same p-value can correspond to a wide range of different Bayes factors, depending on how one models H_1 (i.e., depending on the predictions of the theory being tested). Further, a Bayes factor can distinguish evidence for H_0 (where H_0 is the hypothesis of no effect) from not enough evidence to distinguish H_0 from H_1. A Bayes factor of about 1/3 is the same amount of evidence for H_0 relative to H_1 as is a Bayes factor of 3 for H_1 relative to H_0. A Bayes factor closer to 1 is only evidence "not worth more than bare mention" in Jeffreys' (1939) terms (i.e., not enough evidence to distinguish H_0 from H_1). Thus, a Bayes factor makes a vital distinction not available to those who look only at a p-value (using the H_0 of no effect) – it can distinguish evidence for no effect from not enough evidence to say whether there is an effect. For example, if one obtained a Bayes factor of 0.6, one would not conclude anything about H_1 vs H_0; if such a result corresponded to a non-significant result, that non-significant result does not count against a theory that predicted a difference. This is a fact that should be reflected in the discussion section of a paper – the result should be treated as non-evidential.

The Bayes factor depends on how probable the data are given H_1. Thus, a model is needed of the plausibility of different population values given H_1 (e.g., population mean differences); we call this the model of H_1. The model of H_1 is often called the *prior* of the Bayes factor, but the term "prior" also refers to the prior odds we just referred to. Thus, the word "prior" has two referents in the context of Bayes factors. For example, one might say that "prior odds are purely personal and subjective, and as Bayes factors depend on priors, that makes Bayes factors arbitrary." But this argument conflates the two referents of "prior." Thus, I use two different terms; the term "model of H_1" refers to the mathematical way the predictions of a theory are represented; the "prior odds" refers to the relative probability of the two hypotheses being contrasted. Typically, in results sections of papers (see Dienes, 2021a for examples), I give only the Bayes factor and not the prior odds (though the full Bayesian machinery, including prior odds, can be useful).

Theories, Hypotheses, and Models

Let a substantial theory be a theory that could be tested by the study – a theory for which the possible data could count against. A theory could be, for example, that "cultivating kindness, intensely, to a person who has wronged you, involves suppressing anger and will produce a rebound feeling of anger over the next few hours." From the substantial theory, background assumptions are used to generate predictions. Predictions can be expressed as specific hypotheses with defined independent and dependent variables. For example, one out of several hypotheses relevant to testing the above theory could be: If people practice kindness meditation (with script X) to a person who has wronged them rather than a stranger (the control), a subsequent bump by a confederate as they leave the room will produce more anger-relevant facial muscle activation (from 1 = trace to 5 = maximum).

A model is a probability distribution of effects given a hypothesis. For example, the null hypothesis, H_0, states that there will be no difference between groups in anger on the 1–5 scale. One can represent this as a plot of plausibility (probability density) against different population group differences, with a spike at zero; the only possible difference on this hypothesis is zero (see Figure 23.1) The alternative hypothesis, H_1, states that there is a difference on the anger scale in population means between the two groups. The model of H_1 is the probability distribution for various group differences. What range of differences could be expected by the theory? Are some differences more plausible than others? Note that the theory states that the mechanism is rebound after suppression. Thus, a norming study could be run where one group is asked to consider a person who has wronged them and suppress any anger they feel; the other group does likewise for a stranger. The estimated amount of difference in rebound anger can inform the model of H_1 for the main experiment. For example, the posterior

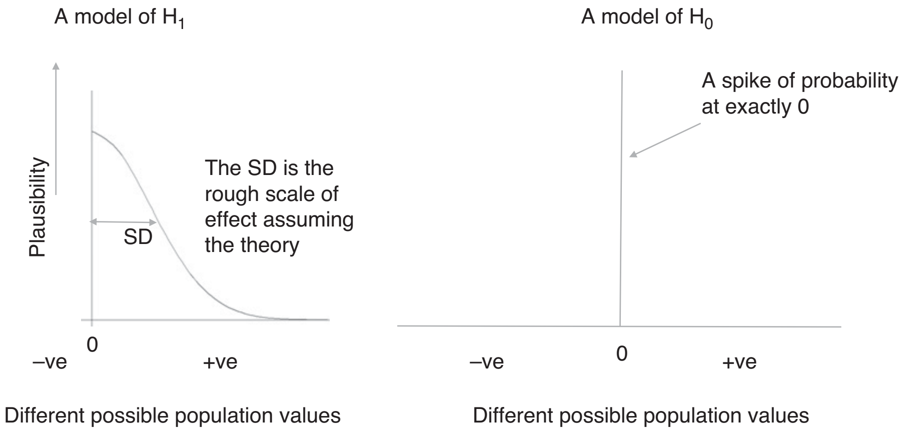


Figure 23.1 *Common models of H_1 and H_0. Each graph is a plot of the plausibility (probability density) of different possible population parameter values (e.g., slopes and mean differences). Let a positive value be in the direction predicted by the theory (e.g., "compassion to enemy" group will show greater anger than "compassion to stranger" group). H_0 is a point H_0 – there is a spike of probability for just one value and no difference between the means (or zero slope). Note that H_1 allows a range of population values consistent with the theory, with smaller values more likely than larger ones.*

distribution of the difference in rebound anger in the norming experiment (i.e., the distribution of the uncertainty in the population mean difference) could be the model of H_1 for the main experiment. Note how theory is used to inform the model of H_1; indeed, how else could we test the theory than testing its actual predictions? Note also that the model of H_1 represents the information we have, to date. Thus, the Bayes factor is in that sense provisional; with better information (e.g., a larger norming study), the model of H_1 could be better informed (see Popper, 1959, p 275).

A Bayes factor indicates the amount of evidence for the model of H_1 relative to the model of H_0. If the background assumptions used in generating the predictions are safe, then evidence counting against H_1 relative to H_0 also counts against the theory that predicts H_1. In this way, the theory itself can be tested. Sometimes different theories lead to the same predictions. In this case, the Bayes factor comparing those theories is comparing the model of H_1 against the same model of H_1, and there is no evidence distinguishing those two theories. For example, one theory could be that the greater the kindness one has felt for a period, the less anger one would feel in the next few hours. The reason the groups in the main study differ in anger, after a bump, is then because they cultivated different degrees of kindness. One could deal with this by asking the two groups to cultivate the same specific degree of kindness – which one would have to assure by finding evidence for H_0 on a measure of kindness. Thus, the evidence by a Bayes factor (B) > 3 (or > k, whatever was good enough) for the H_1 (predicted by the substantial theory) versus H_0 supports the substantial theory relative to other theories that predict H_0. Or, more generally, a $B > 3$ for H_1 versus H_2 (predicted by another theory) supports the theory predicting H_1 over the theory predicting H_2.

The evidence provided by a Bayes factor comparing H_1 with H_0 supports the substantial theory over others that predict H_0, but only to the extent that the model of H_1 represents the predictions of the theory. The theory should do work in generating those predictions; without the theory, ideally other background knowledge would not make the same predictions (Popper, 1963). In this example, the theory was itself used in generating predictions; the model of H_1 was based on a norming experiment estimating rebound anger because that is the mechanism that the theory specifically postulates. The rebound norming experiment would not be relevant to a theory that postulated a difference in anger in the main study because the groups differed in amount of kindness cultivated. For example, the predictions about anger after a bump may turn out much the same in a particular study for the rebound theory and the kindness theory, but that would have to be shown.

That is, one should not think there are two entirely separate phases to theory testing – working out the predictions of a theory and statistical hypothesis testing without reference to theory. One could not get evidence for no effect unless there was a theory to specify the range of effects predicted. There is no such thing as evidence for nothing being there, independent of what size effect there could be. One cannot test the absence of something in the absence of theory. Conversely, the evidence for nothing being there can be different relative to different theories. A theory that says the thing looked for can only be very small (is there a flea bite on the arm?) will need a smaller standard error to get evidence for it not being there compared with a theory that claims it could be very large (is there a dog bite on the arm?). The theory does not have to be mathematical or computational to make relevant predictions; psychological theories predict more than researchers typically realize.

Just as several theories can make the same or similar predictions, so the prediction of a single theory can be modeled in several ways, each model just as reasonable a representation of the prediction as the others. What is needed is to show robustness of the conclusions over different models of H_1 that are roughly equally reasonable as representations of the prediction. I will discuss a simple way of doing this in the next section.

How to Model H_1

Predictions can often be reasonably fixed by scientific context. Some simplifying assumptions are also useful. The precise shape of the model of H_1 is rarely relevant to a theory. A simple approach is to use a distribution with a mode of zero (e.g., a normal distribution – presuming zero is the value predicted by H_0). This means the shape puts most probability around the same value as H_0; thus, the shape can make it slightly harder to discriminate H_1 from H_0. That is, when the Bayes factor does give good evidence for the one hypothesis rather than the other, it is despite rather than because of the precise model of H_1. If the theory predicts a direction of the effect, the distribution below zero can be removed, leaving a half-normal distribution (see Figure 23.1). The standard deviation (SD) of the normal or half-normal distribution scales the rate at which the curve drops toward zero. Thus, one sets the standard deviation to the approximate scale of effect predicted. Only 5% of the area of distribution is beyond two standard deviations out; thus, twice the scale of effect is also the rough maximum. Another way of putting this is, if one has information to specify a rough maximum, set the standard deviation of a half-normal to half that maximum.

A rough maximum can often be estimated by determining the possible "room to move" (Dienes, 2019). Klaschinski et al. (2017) investigated whether briefly assuming an expansive rather than contracting pose would increase one's performance in a subsequent interview. Participants were interviewed, and raters evaluated how well the subject performed on a 1 ("awful") to 7 ("amazing") scale. The theory tested was that assuming a posture typical of being confident rather than insecure would make one perform better in a demanding situation. People in the power posing condition were rated on average 4.23 Likert units; those in the control condition were 4.17 Likert units. The mean difference is 0.06 Likert units (SE = 0.21 Likert units) and is non-significant at the 5% level (t =0.06/0.21 = 0.28, p = 0.78). A non-significant difference by itself allows no conclusion about whether there is a population difference. To see which way the evidence points, the possible size of the population differences needs to be determined. The control condition had a mean of about 4; thus, the room to move – the biggest difference if the theory were true – would be if the power pose group scored the maximum on the scale (7), giving a maximum possible difference of (7–4) = 3 Likert units.

Given this estimated maximum, the plausibility of different possible population differences, if the theory were true, can be modeled as a half-normal with a mode of zero and an SD of the maximum divided by 2 = 3/2 = 1.5 Likert units. The "half" of the half-normal indicates the theory predicts in a certain direction; performance is predicted to be higher in the expansive rather than contracting pose condition. The SD of the half-normal indicates the rough scale of effect expected (1.5 Likert units). The Bayes factor can be calculated and written as $B_{\mathrm{HN}(0,1.5)} = 0.18$, where the "HN" indicates a half-normal was used to model H_1, the "0" refers to the mode and the "1.5" to its SD. To obtain this Bayes factor, go to www.bayesfactor.info for a ShinyApp. Enter "0.06" for the "mean difference," "0.21" for the "standard error," "1.5" for the "hypothesized mean difference," and click "positive 1-tailed" to make it half-normal. Here, the work done by the theory is in indicating the direction of the effect; the constraints in other aspects of the data themselves then limit how big the effect could be. The value of 0.18 is conventionally considered as moderate evidence for no effect, as it is less than 1/3.

In fact, in this case, there is more information. The work carried out by Klaschinski et al. (2017) was a replication of that of Cuddy et al. (2015). In the original study, Cuddy et al. found that the power posing group were rated as 4.63 Likert units; the control group 3.81 units. Thus, the difference was 0.8 Likert units with a standard error of 0.28 Likert units. Using 0.8 Likert units as a scale factor (instead of the previous 1.5), $B_{\mathrm{HN}(0,0.8)} = 0.32$. Since this is less than a third, the same qualitative conclusion results – moderate evidence for no effect. In fact, for this second Bayes factor, the theory is slightly different. Using the effect from a previous study to inform a replication attempt tests the theory that the methods given in the method section of the original study lead to the sort of results given in the results section (this is called the replication hypothesis; see Popper, 1959, p. 66).

As the original study was an attempt to test the same substantial theory, the test of the replication hypothesis is relevant to the substantial theory (for interpreting evidence against a replication hypothesis, see Edlund et al., 2021). In any case, the use of 1.5 for the scale factor in the first Bayes factor is somewhat arbitrary. How robust is the conclusion to different scale factors? One can report a "robustness region" – the set of scale factors that lead to the same qualitative conclusion (e.g., $B < 1/3$; Dienes, 2019); in this case, it can be notated $\mathrm{RR}_{B<1/3}$ [0.8, >6]. The "0.8" indicates that 0.8 is the lowest the scale factor can be and there is still moderate evidence for H_0; the scale factor can exceed the length of the scale (1–7), and there would still be evidence for H_0. The lowest scale factor implies the effect could plausibly be between 0 and 1.6 Likert units (i.e., from 0 to twice the scale factor). If there is no compelling reason why the effect must be more constrained than this, the conclusion is robust.

How does one construct a model of H_1? One uses assumptions that are either simple or scientifically motivated to derive predictions. The predictions can then be tested against the observed effect. One might be tempted to think "I do not know what effect sizes my theory predicts; the observed effect is the best information I have about its size, so why don't I use this to model H_1?" Then, the predictions can never clash with the data, so the data can never count against the theory. To test the theory, we need the model of H_1 to possibly clash with the data. In the power posing example, if we used the observed effect of 0.06 Likert units as the scale factor, we obtain $B_{\mathrm{HN}(0,\, 0.06)} = 1.02$; this does not count against the theory. When we use the room-to-move heuristic, by contrast, even though we are drawing on an aspect of the same data, this does not compromise the ability of the test to find evidence against the theory – as indeed it did in this example (Dienes, 2019; see also Devezer et al., 2020).

Severe Testing

Popper (1963) defined a severe test as one in which, if the theory is false, there is a high probability the theory will be found false. Popper (p. 526) formalized the notion of a severe test by representing it as the ratio of the probability of the predicted outcome given the theory to the probability of the outcome assuming the theory were false (and assuming the rest of background knowledge). That is, a severe test is one that can generate an extreme Bayes factor (i.e., one that is very large or very small) depending on whether the theory is true or false. Bayes factors, as a conceptual tool, therefore go hand in hand with determining the conditions for a severe test (van Dongen et al., 2020; Vanpaemel, 2020).

Assume that all modeling assumptions are satisfactory (e.g., the distribution of the data). If H_0 is true, a Bayes factor is, with increasing subjects, eventually driven to 0. That is, if a theory is false (H_0 is true), with enough data, a Bayes factor is driven to show the theory false. Likewise, if there is a difference, then the Bayes factor is, with increasing subjects, eventually driven to infinity. That is, if H_0 is false, with enough data, H_0 is eventually found false, but this only occurs if our models represent everything relevant in an approximately good enough way. One crucial aspect of the modeling is how subjects' data are generated by the world for a given population mean difference (or value of whichever parameter is being tested). If a sample mean difference of 0.8 is found, then how plausible is this mean for different possible population mean differences? This is the likelihood function.

This function does not take on a simple form when researchers cherry pick and hack. For example, assume the researcher determines how the sample mean difference can be made as close to a predicted value as possible, by playing with different ways of excluding outliers, removing participants according to different equally sensible exclusion criteria, adding or taking away covariates, and so on. "*B*-hacking" is this process of trying different analyses to push *B* in the direction one wants. With this new likelihood function, the probability of obtaining the "data" – represented as the finally cited mean difference – may be very similar given H_0 as given H_1.

Consider a case, after some *B*-hacking, where an uncorrected Bayes factor (i.e., one that does not consider hacking) shows evidence for H_1, even when H_0 is true. The Bayes factor, using the appropriate likelihood function (i.e., considers hacking, how cited mean differences are generated given a population parameter value), should be close to 1, and the whole process thereby is uninformative. The test would in no way be severe, as a false theory would not often be found false. Thus, the concept of a Bayes factor, by measuring evidence, shows why cherry picking and hacking render data non-evidential. The "data" are just about as probable given H_0 as H_1.

Cherry Picking

To get the appropriate Bayes factor, hacking and cherry picking must be considered, as they influence the likelihood function. Modeling the effects of hacking may be difficult in a real-world case. So, rather than try to model the effects of generic hacking, it is better to set up safeguards to protect against their effect or, in the absence of safeguards, to acknowledge the potential role of *B*-hacking. In terms of safeguards, one can preregister the analytic protocol, including the modeling of H_1, or use a full-fledged registered report (Chambers, 2019; Dienes, 2021b). Alternatively, one can use blind analyses, in which the analyst is given several data sets and does not know which is the real one (MacCoun & Perlmutter, 2015). One may be able to rely on simplicity and strong theoretical arguments (Szollosi et al., 2020), though, given the ease of hacking in even simple situations, this is unlikely to be a complete solution (Wagenmakers, 2019). In terms of acknowledging the issue, one may treat the result as more tentative than the Bayes factor indicates at face value and then seek to replicate it; the original result could, for example, be published as an exploratory report (McIntosh, 2017).

Stopping Rule

The formula for the Bayes factor given at the beginning of this chapter showed that is a complete summary of the evidence – how much one should change one's beliefs, assuming that the axioms of probability represent idealized constraints on the strength of belief. Thus, all that matters for evidence is the probability of the data, given the hypotheses contrasted. In particular, it is not relevant to interpreting a Bayes factor as evidence when one stops collecting data. That is, one need not state, in advance, a stopping rule when using Bayes factors (Dienes, 2016; Hendriksen et al., 2020; Rouder & Haff, 2020). Optional stopping, a questionable research practice for significance testing, is not one for Bayes factors. In this way, the evidence indicated by Bayes factors has the same property as evidence does in all other arenas, apart from significance testers doing research (Wagenmakers et al., 2019). In no other case is evidence adjusted according to a stopping rule, be it detectives searching for evidence for the murderer, children learning to speak, animals learning where the best food is, or significance testers learning a foreign language in their spare time. Optional stopping with Bayes factors (e.g., stopping when the Bayes factor is greater than 10 or less than 1/10 to ensure strong evidence) allows severe testing of either H_1 or H_0.

Multiple Testing

Multiple testing of different hypotheses, all relevant to a theory, also shows the way in which Bayes factors reflect severe testing requirements. Consider the substantial theory that extrinsic motivation reduces intrinsic motivation because, if people see that they behave a certain way to get an extrinsic reward like money, they conclude they do not want to behave that way for its own sake. An experiment puts children in a room and rewards them for playing with some toys rather than others – randomly selected for different groups. The children are then observed surreptitiously for how much time they play with each toy when they think they are alone. It is found that children play less with the rewarded than the control toys but only for females in the afternoon. The Bayes factor for that specific contrast is greater than 3. The authors conclude that the theory is supported, perhaps especially for people more sensitive to the sorts of rewards used. This is a case of multiple testing.

Multiple testing can result in increasingly implausible hypotheses being tested, including hypotheses that do not follow simply from the substantial theory. The test of the substantial theory is the overall difference in time playing with rewarded rather than control toys. All data relevant to the theory must be included in evaluating the theory. Researchers may be vaguely aware that significant results carry more evidential value than non-significant ones, so they may think they can focus on significant results and safely ignore the non-significant results. However, cherry picking is wrong in any school of statistical inference. From a Bayesian perspective, it is apparent that all data relevant to testing the theory provide some degree of evidence. Thus, the test of the theory is provided by all the data (see Dienes, 2008, pp. 108–114, and Dienes, 2016, for different examples). If the theory, as stated, is wrong, the data as a whole, reflecting the simplest test of the theory, are likely to show it wrong, given enough data.

In this example, the Bayes factor for the main effect of reward may be less than 1/3, in which case the evidence counts against the theory. There is also evidence for a peculiar combination of conditions leading to a reward effect, but this cannot be used in isolation as support for the theory. What often makes the results of multiple testing seem unsatisfactory is the low prior probability of the hypotheses that get evidence, given the other hypotheses that had evidence against them. Alternatively, multiple testing often finds support for hypotheses that do not follow in a simple way from a simple theory. One might explicitly consider the prior odds of the different hypotheses and thereby demand more evidence (i.e., a higher Bayes factor threshold). In the example considered here, the simplest approach, rather than adjusting thresholds, may be to acknowledge that the substantial theory has evidence against it and leave it at that. Alternatively, one could conjecture a simple and bold theory that might explain why females in the afternoon, and not other combinations, show the effect, derive predictions from such a theory, and test them in a further study.

Notice that no general Bonferroni correction (or other familywise correction procedures) quite gets the point right in dealing with multiple comparisons. The Bonferroni correction used by significance testers says one should use a significance threshold of $0.05/k$ if one conducts k tests in a family of tests. What is a family of tests (see Kruschke, 2011)? A temptation might be to say that a family is all the tests relevant to a theory, but theories come in hierarchies; there is the most general theory (e.g., dissonance theory), there is an application of that theory to a particular problem (intrinsic motivation), and there are more specific applications (children playing with toys). There is no single answer to what k should be for theory testing because there are simultaneously several theories at play (Dienes, 2016). Imagine a paper that looked at children with toys and adults with artwork, but significance testing demands a single once-and-for-all answer. A Bayes factor is a measure of evidence relative to theories – if you want to know the evidence for theory X, consider all data relevant to theory X. One can simultaneously consider the evidence for different theories in a hierarchy with different Bayes factors. Because significance testing considers that tests must be controlled relative to a person, it fails to be able to measure evidence relative to theories.

Consider a theory tested by asking five questions on a Likert scale. Each question alone may provide some evidence for the theory (e.g., $B > 3$ and $p < 0.05$ for each question). With a Bonferroni correction (requiring $p < 0.05/5$, or $B > 3 \times 5$), each question may fail to provide good enough evidence for the theory. However, when the questions are combined together in a single measure, the evidence would be $B > 3$, $p < 0.05$. That is, forming an overall conclusion by applying Bonferroni to the individual questions is inappropriate in this case; the automatic application of familywise error correction does not solve the problem of multiple testing.

Where the prior probability of one hypothesis being true can be roughly assessed by the number of hypotheses tested, a Bonferroni correction can be useful (see Westfall et al., 1997). For example, when looking at 20,000 different genes to see if any correlates with a phenotype, it makes sense to take into account how improbable each H_1 is – approximately 1/20,000. In the context of such implausible individual hypotheses, a more severe than typical test is called for – a Bayes factor threshold of 3 × 20,000 can be used. In general, when data dredging, one should consider prior probability because the whole point of dredging is that most hypotheses are wrong.

So How Can There Be Evidence for Theories That Go Beyond the Data?

Scientific theories are interesting because they are bold and go beyond the data (Popper, 1963, p. 330). After finding that his subjects were slower at naming colors when the words were incongruent, but were not slower at naming words when the colors were incongruent, Stroop (1935) did not say "My theory is that the students at George Peabody College for Teachers in 1934 . . . "; instead he said that, in general, "the associations that have been formed between the word stimuli and the reading response are evidently more effective than those that have been formed between the color stimuli and the naming response" (Stroop, 1935, p. 660). According to Popper, he was not inducing the claim (and hence having to stay close to the data) but rather testing it. As the claim is meant to be general, if it fails for students at George Peabody College for Teachers, it fails; Stroop's study constituted a test of the claim. The more general the claim, the more opportunities there are for finding Bayes factors that go against it. Hence, bold general claims are easier to severely test; more situations can constitute tests of general claims than claims that "stay close to the data."

When testing a theory, one conjectures a world in which the theory may be true or false. There is a probability structure that arises because of the way that conjectured world generates data. Within that world, probabilities change in the light of data. In that sense, contrary at face value to Popper's claims that induction does not exist, induction can occur in the sense of changing probabilities of theories in the light of data (Jeffreys, 1939). That is what evidence is – something that normatively changes strength of belief. But Popper is right. The induction does not occur in any absolute sense; it occurs in the conjectured world. In that sense, every statistical test is a thought experiment (Greenland, 2017). It is a conjectured world, but part of the conjecture is that the conjectured world is a good enough approximation of the real world, so that conclusions in the conjectural world are relevant to the real world. Ultimately, the only use of the conjectural world is to test claims about the real world.

How can we check if the conjectured world is relevant to the real world? The only option is to test its assumptions (Morey et al., 2013; Notturno, 1999). We embed the conjectured world in a larger also-conjectured world (Kruschke, 2013a). For example, the larger world may assume different degrees of skew in the likelihood distribution while the world in which we conducted our test may assume a normal likelihood; both worlds assume a certain family of distributions. In this way, each assumption can, in principle, be tested one by one. Significance tests are often used by researchers for checking assumptions, yet a significant violation does not mean that the violation was extreme enough to make any difference to conclusions; a non-significant result does not mean there was not a violation that makes a huge difference to conclusions (Krushcke, 2013a). That is, as in any area of testing, to test if there is nothing there, one must consider how big the thing is that is relevant. Was the violation large enough to alter conclusions?

More work is needed on how well calibrated Bayes factors remain depending on degrees of violation of assumptions (Rouder & Haaf, 2020). One rule of thumb is that the same conditions that lead to error rates changing markedly for significance tests are likely to be the same that lead to the equivalent Bayes factors to be badly calibrated. For example, in conditions where one should use adjusted degrees of freedom for a *t*-test, because of variance inequality, one should use those same adjusted degrees of freedom for the corresponding Bayes factor. This is because, if one makes simplifying assumptions, Bayes factors become monotonic with *p*-values (Benjamin et al., 2018); thus, whatever conditions influence the correctness of those *p*-values will influence the corresponding Bayes factors. If those Bayes factors are so affected, so are the Bayes factors one would use in real situations. Piecemeal testing of assumptions (Mayo, 2018) can help determine where one's conjectured world ceases to be relevant to the real world.

In deriving predictions from a theory, and thereby setting up a conjectural world, the probabilities in that world (e.g., in the form of the model of H_1) are idealized subjective probabilities – strengths of belief assigned to an idealized person confronting the relevant background knowledge. Once postulated, the model becomes objective in the sense that its consequences can be discovered and the assumptions criticized by anyone (Popper, 1972). A model of H_1 is meant to follow from theories using simple and otherwise well-tested assumptions; what matters are the reasons why the assumptions are made, not what any specific individual believes. In that sense, one need not follow a purely subjective Bayesian approach that treats predictions (as models of H_1) as purely subjective feelings of what size an effect may be. One does need to make a judgment that the model is satisfactory enough to use it, but any judgment against an assumption is simply a promissory note that reasons can be found for criticizing that assumption; what will matter in the end are those reasons that can be made public and criticized by anyone else (Miller, 1999; Notturno, 1999). Similarly, one need not follow a purely objective Bayesian approach that uses default models of H_1 – a single model of H_1 for any theory (e.g., a Cauchy with scale factor 1, as used by Rouder et al., 2009). Testing theories means representing the predictions of specifically those theories and confronting those predictions with data.

Criticisms of Bayes Factors

1. Different Models of H_1 Give Different Answers

This is sometimes called the problem of prior sensitivity; if you change the model of H_1, you get different Bayes factors – so, how can you know what the evidence is? (Kruschke, 2013b). However, this is not a fault of Bayes factors but a virtue.

Different theories, or different assumptions connecting theory to predictions, make different predictions. It is useful and necessary that different predictions lead to different degrees of evidence for different theories. That is, the model of H_1 needs to represent the predictions of the theory put to test. The problem then becomes one of knowing what your theory predicts. Dienes (2019, 2021a) discusses different heuristics for modeling H_1 given different theories and scientific contexts. Inferential ambiguity can arise when the same prediction can be equally well modeled in different ways; one way of partially addressing this is with the robustness region mentioned above (Dienes, 2019). Also, as mentioned above, prior sensitivity provides extra analytical flexibility and opportunities for *B*-hacking. This is partly addressed by robustness regions and should also be addressed by other methods for dealing with analytical flexibility (e.g., preregistration; see previous discussion; Dienes, 2021b).

2. A Default Bayes Factor Is Not Relevant to Your Theory

Tendeiro and Kiers (2019) pointed that default Bayes factors lack clear empirical justification for any specific application. Make sure the model of H_1 represents the predictions of your theory. Always explicitly state what your model of H_1 is and give an objective reason why you set any parameter value. Do not use a default scale factor just because it is already there in the software; treat it as an invitation for you to consider whether the default settings are relevant. The urge to use defaults may be based on regarding inferential statistics as being insulated from theory – whatever the theory, one conducts the standard inferential statistics, and they indicate whether there is an effect there or not. On this insulating approach, conclusions about whether there is an effect can then be used to count for or against theoretical predictions, no matter what the theory is. This simplistic separation of statistics from theory may be one reason why we have been having problems properly testing our theories.

3. The Point H_0 Is Never True; Therefore, the Only Thing to Do Is to Estimate

Meehl (1967) argued that, in the case of real-world correlations, all point H_0s are false (i.e., the H_0 of there being exactly no difference or no slope); surely, all point H_0s are false to some decimal point for any hypothesis in the social and behavioral sciences if only because modeling assumptions are never exactly true. This claim can be used to argue for the futility of performing any hypothesis test against the point H_0, whether using significance testing or Bayes factors. On this view, one should always just reject the point H_0 and be done with it (see Baguley, 2012, p. 368; Wagenmakers, 2017, for arguments and replies).

Bayes factors can compare any two models. The point H_0 is often used as one of the models to test the claim that there is something there rather than nothing. Baguley (2012) illustrates how accurate a point H_0 can be with Wiseman and Greening's (2002) online experiment with 27,856 participants, testing the existence of extrasensory perception (ESP) with a chance baseline of 50%. The 95% confidence interval (CI) was [49.6%, 50.2%]. That is, even though the statistical assumptions that lead to an estimated predicted 50% for no ESP must be approximate, whatever the true H_0 is, it lies within an interval of roughly [−0.3%, + 0.3%] around 50%. That means we can trust the same modeling assumptions in other similar contexts to be accurate within this interval. Now, imagine the same modeling assumptions with a similar experiment, investigating a more mundane effect, with a roughly expected effect size of about 5% above the baseline. Further, one might decide that a minimally interesting effect would be any value 0.5% above the baseline. Results just give support for H_1 using a point H_0: 55% correct standard error is 2.5%, $B_{H(0,5)} = 4.27$ against a point H_0. A different online Bayes factor calculator is illustrated this time to obtain the Bayes factor, as this calculator allows interval H_0s. Go to https://bayesplay.colling.net.nz/. For likelihood, click on "normal"; enter "5" for mean, and "2.5" for SD. For alternative prior, click on "normal". Enter "0" for mean and "5" for SD. To make it a half-normal, click on "lower limit" and enter "0." For null prior, enter "point." Enter "0" for point. Click "calculate."

A Bayes factor compares any two models, so we can compare a similar model of H_1 (but with a mode of 0.5 above 50%) against an interval H_0 around 50% [−0.5%, +0.5%]. This gives $B_{H(0.5,5)} = 4.44$ – virtually the same answer. To obtain this Bayes factor, proceed in the same way as for the previous paragraph; go to null prior, click "uniform" and enter "−0.5" for the lower limit and "0.5" for the upper limit. Finally, for the lower limit for H_1, change to "0.5." Click calculate. In sum, the point H_0 is a perfectly adequate approximation of the true interval H_0. This will be generally true whenever the standard error is large compared to the null interval. The point H_0 will often be a perfectly adequate approximation in well-controlled experimental research; this is similar to assuming any particular distribution, or any particular statistical model at all, is a useful approximation in any research. If we are going to reject all point H_0s because they are not exactly true, we must similarly reject all models whatsoever. When big data are involved, the standard error may be very small; then, it will be important to be clear about how H_0 should be modeled – maybe with a uniform distribution – and the plausibility of H_1 may be zero below a minimally interesting value (see Palfi & Dienes, 2019 for Bayes factors with interval H_0s and an example case; see Skora et al., 2020 for another type of example using interval H_0s).

Hypothesis testing is needed whenever a researcher wants to ask whether something exists: Should a term be in the model? Is there an interaction? Are people performing at a chance baseline? If the question of something existing can be taken for granted, then all one needs to do is estimate. However, if one only estimates, one cannot infer from the estimation that there is no effect nor treat the estimation as grounds for asserting that something exists rather than not existing at all (existence has been presumed by estimation, not tested). With these provisos, a paper with only estimation may be a useful option. Can you draw all the conclusions you want from saying that, whatever the effect is, it is plausibly between such and such bounds? It is an option I have followed for several papers (e.g., Palfi et al., 2020).

4. A Bayes Factor Does Not Control Error Rates

Mayo (2018) criticized Bayes factors for not controlling error rates. A Bayes factor measures the strength of evidence; what you do with evidence (over many cases) determines error rates. Fixing one (evidence or error rates) at a desired value does not fix the other at any particular value. The intuition behind criticizing Bayes factors for not precisely controlling error rates is presumably that given errors rates are not fixed, and one should be more cautious in updating beliefs from Bayes factors. In sum, the critic says: Don't take Bayes factors seriously. But this argument is back to front. Evidence, not error rates, is how much you should change your confidence in H_1 versus H_0. The critic's intuition presumably arises because the more evidence one has in the long run, the lower error rates tend to be; one may mistakenly feel that, unless error rates are precisely fixed, evidence cannot be "true" evidence.

Evidence constrains error rates without fixing them. In fact, Bayes factors can minimize the weighted sum of Type I and II errors (see DeGroot, 1986, p. 444 for likelihood ratios; Pericchia & Pereira, 2016 for Bayes factors). Evidence and error rates are related, and strong evidence will be associated with small (but not fixed) error rates (e.g., see Tables 1 and 2 in Dienes, 2016). Still, if one's interest is in evidence, it is Bayes factors that measure it. Error rates may be useful in understanding what a certain quantity of evidence means (see Hendriksen et al., 2020); if error rates seem too high, then that level of evidence may be too low for your needs. It is instructive to consider how, with evidence less than infinite, there must always be a probability that the evidence reverses as more data come in; with evidence as low $B = 3$ (roughly, $p = 0.05$), reversals are reasonably common (see Tables 1 and 2 in Dienes, 2016).

In sum, there are no convincing arguments against Bayes factors as such, only against their misuse or misunderstanding.

Conclusion

I personally follow a policy of a "*B* for every *p*." When hypothesis testing, I report both *p*-values and Bayes factors, whose model of H_1 is informed by the theories tested. Inferences are always with respect to the Bayes factors, but *p*-values are given so people can see their relation to Bayes factors. In fact, there is a Bayesian interpretation of *p*-values. Imagine we are using half-normal models of H_1 in a paper. If a result is significant, $p < 0.05$, then there is some standard deviation for the half-normal for which $B > 3$; likewise, if $p < 0.01$, there is some standard deviation for which $B > 10$. That standard deviation may not be relevant for the theory, or it may not be the only relevant one; only the Bayes factors with scientifically motivated standard deviations will measure evidence relevant to the theory. The *p*-value still carries information, so it is useful to quote it.

If one decides a threshold for regarding evidence as adequate (e.g., 3), then one can say that when $B > 3$, "There was evidence for a main effect of . . . "; when $B < 1/3$, "There was evidence for no main effect . . . "; and when *B* is between those values (i.e., close to 1), “There was no evidence one way or the other for a main effect . . ..” In the latter case, one would draw no theoretical conclusions in the discussion about whether there is an effect (see Dienes, 2021a for example). However, when $B < 1/3$, one could use that result to count against a theory that predicted the H_1 tested. That is, theories can actually be tested.

Bayes factors are also useful when more participants are needed. If one has already collected data, and a reviewer asks for more, significance testing is now illegitimate. For the same reason, combining studies together for significance testing is also illegitimate. The stopping rule must be respected for significance testing. Thus, evidence is wasted. How can a theory be severely tested if the data fall short of severely testing them and it is forbidden to collect more? Because Bayes factors measure evidence, and one can always accumulate evidence until one has enough, Bayes factors must be used for hypothesis testing once data are to be combined.

In sum, Bayes factors are an invaluable practical and conceptual tool for understanding, and practically engaging in, the severe testing of theories.

References

Baguley, T. S. (2012). *Serious Stats: A Guide to Advanced Statistics for the Behavioral Sciences*. Palgrave Macmillan.

Benjamin, D. J., Berger, J. O., Johannesson, M., et al. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Chambers, C. D. (2019). What’s next for registered reports? *Nature*, 573(7773), 187–189.

Cortex (2021). Guidelines for users. Available at: http://cdn.elsevier.com/promis_misc/PROMIS%20pub_idt_CORTEX%20Guidelines_RR_29_04_2013.pdf.

Cowles, M. & Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist*, *37*, 553–558.

Cuddy, A. C., Wilmuth, C. A., Yap, A. J., & Carney, D. R. (2015). Preparatory power posing affects nonverbal presence and job interview performance. *Journal of Applied Psychology*, *100*, 1286–1295.

DeGroot, M. H. (1986). *Probability and Statistics*, 2nd ed. Addison-Wesley.

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Buzbas, E. O. (2020). The case for formal methodology in scientific reform. *bioRxiv.* https://doi.org/10.1101/2020.04.26.048306

Dienes, Z. (2008). *Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference*. Palgrave Macmillan.

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, *72*, 78–89.

Dienes, Z. (2019). How do I know what my theory predicts? *Advances in Methods and Practices in Psychological Science*, *2*, 364–377.

Dienes, Z. (2021a). How to use and report Bayesian hypothesis tests. *Psychology of Consciousness: Theory, Research, and Practice*, *8*, 9–26

Dienes, Z. (2021b). The inner workings of registered reports [Preprint]. *PsyArXiv.* https://doi.org/10.31234/osf.io/yhp2a

Dienes, Z. & McLatchie, N. (2018). Four reasons to prefer Bayesian over significance testing. *Psychonomic Bulletin & Review*, *25*, 207–218.

Edlund, J., Cuccolo, K., Irgens, M. S., Wagge, J. R., & Zlokovich, M. S. (2021). Saving science through replication studies. *Perspectives on Psychological Science*, March 8. https://doi.org/10.1177/1745691620984385

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd.

Greenland, S. (2017). The need for cognitive science in methodology. *American Journal of Epidemiology*, *186*, 639–645.

Hendriksen, A., de Heide, R., & Grünwald. P. (2020). Optional stopping with Bayes factors. Available at https://arxiv.org/pdf/1807.09077.pdf.

Jeffreys, H. (1939). *The Theory of Probability*. Oxford University Press.

Klaschinski, L., Schnabel, K., & Schröder-Abé, M. (2017) Benefits of power posing: Effects on dominance and social sensitivity, *Comprehensive Results in Social Psychology*, *2*, 55–67.

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*, 299–312.

Kruschke, J. K. (2013a). Posterior predictive checks can and should be Bayesian. *British Journal of Mathematical and Statistical Psychology*, *66*, 45–56.

Kruschke, J. K. (2013b). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General*, *142*, 573–603

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, *44*, 187–192.

Lindley, D. V. (2014). *Understanding Uncertainty*, revised edition. John Wiley & Sons.

MacCoun, R. & Perlmutter, S. (2015). Hide results to seek the truth. *Nature*, *526*, 187–189.

Mayo, D. G. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge University Press.

Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*, 103–115.

McIntosh, R. D. (2017). Exploratory reports: A new article type for Cortex. *Cortetx*, *96*, A1–A4.

McPhetres, J., Albayrak-Aydemir, N., Barbosa Mendes, A., et al. (2021). A decade of theory as reflected in psychological science (2009–2019). *PLOS One*, March 5. https://doi.org/10.1371/journal.pone.0247986

Miller, D. (1999). *Critical Rationalism: A Restatement and Defence*. Open Court.

Morey, R. (2018). Redefining statistical significance: The statistical arguments. Available at: https://medium.com/@richarddmorey/redefining-statistical-significance-the-statistical-arguments-ae9007bc1f91.

Morey, R. D., Romeijn J. W., & Rouder J. N. (2013). The humble Bayesian: Model checking from a fully Bayesian perspective. *British Journal of Mathematical and Statistical Psychology*. *66*, 68–75

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, *72*, 6–18.

Notturno, M. A. (1999). *Science and the Open Society*. Central European University Press.

Palfi, B. & Dienes, Z. (2019). When and how to calculate the Bayes factor with an interval null hypothesis. *PsyArXiv*. https://doi.org/10.31234/osf.io/9chmw

Palfi, B., Moga, G., Lush, P., Scott, R. B., & Dienes, Z. (2020). Can hypnotic suggestibility be measured online? *Psychological Research*, *84*, 1460–1471. https://doi.org/10.1007/s00426-019-01162-w

Pericchia, L. & Pereira, C. (2016). Adaptative significance levels using optimal decision rules. *Brazilian Journal of Probability and Statistics*, *30*, 70–90.

Popper, K. R. (1959). *The Logic of Scientific Discovery*. Hutchinson.

Popper, K. R. (1963). *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge.

Popper, K. R. (1972). *Objective Knowledge: An Evolutionary Approach*. Oxford University Press.

Rouder, J. & Haaf, J. M. (2020). Optional stopping and the interpretation of the Bayes factor. https://doi.org/10.31234/osf.io/m6dhwR

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.

Skora, L., Livermore, J. J. A., Dienes, Z., Seth, A., & Scott, R. B. (2020). Feasibility of unconscious instrumental conditioning: A registered replication. *PsyArXiv*. https://doi.org/10.31234/osf.io/p9dgn

Stigler, S. M. (1999). *Statistics on the Table: The History of Statistical Concepts and Methods*. Harvard University Press.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643–662

Szollosi, A., Kellen, D., Navarro, D. J., et al. (2020). Is preregistration worthwhile? *Trends in Cognitive Sciences*, *24*, 94–95.

Tendeiro, J. N. & Kiers, H. A. L. (2019). A review of issues about null hypothesis Bayesian testing. *Psychological Methods*, *24*(6), 774–795.

van Dongen, N. N. N., Wagenmakers, E., & Sprenger, J. (2020). A Bayesian perspective on severity: Risky predictions and specific hypotheses. *PsyArXiv*. https://doi.org/10.31234/osf.io/4et65

Vanpaemel, W. (2020). Strong theory testing using the prior predictive and the data prior. *Psychological Review*, *127*, 136–145, http://dx.doi.org/10.1037/rev0000167

Wagenmakers, E. (2017). How to test interval-null hypotheses in JASP. Available at: https://jasp-stats.org/2017/10/25/test-interval-null-hypotheses-jasp/.

Wagenmakers, E. (2019). A breakdown of "preregistration is redundant, at best". Available at: www.bayesianspectacles.org/a-breakdown-of-preregistration-is-redundant-at-best.

Wagenmakers, E., Gronau, Q. F., & Vandekerckhove, J. (2019). Five Bayesian intuitions for the stopping rule principle. *PsyArXiv*. https://doi.org/10.31234/osf.io/5ntkd

Westfall, P. H., Johnson, W. O., & Utts, J. M. (1997). A Bayesian perspective on the Bonferroni adjustment. *Biometrika*, *84*, 419–427.

Wiseman, R. & Greening, E. (2002) The mind machine: A mass participation experiment into the possible existence of extrasensory perception. *British Journal of Psychology*, *93*, 487–99.

24 Introduction to Exploratory Factor Analysis: An Applied Approach

Martin Sellbom and David Goretzko

Abstract

This chapter provides an overview of exploratory factor analysis (EFA) from an applied perspective. We start with a discussion of general issues and applications, including definitions of EFA and the underlying common factors model. We briefly cover history and general applications. The most substantive part of the chapter focuses on six steps of EFA. More specifically, we consider variable (or indicator) selection (Step 1), computing the variance–covariance matrix (Step 2), factor-extraction methods (Step 3), factor-retention procedures (Step 4), factor-rotation methods (Step 5), and interpretation (Step 6). We include a data analysis example throughout (with example code for R), with full details in an online supplement. We hope the chapter will provide helpful guidance to applied researchers in the social and behavioral sciences.

Keywords: Exploratory Factor Analysis; Factor Analysis; Internal Structure; Measurement Modeling; Latent Variable Modeling

Introduction

Exploratory factor analysis (EFA) has been an incredibly popular statistical technique in the social and behavioral sciences for over a century (e.g., Goretzko et al., 2021). In the first author's field of psychological assessment, for instance, EFA has been used to elaborate on the internal structure of psychological tests since its inception, and it remains popular today. Open any issue of *Psychological Assessment*, *Journal of Personality Assessment*, or *Organizational Research Methods*, to mention just a few, and you are bound to see articles that used EFA as a method to develop and/or evaluate operationalizations of various constructs. Although popular in psychology, the importance of EFA has been identified in many other areas of the social and behavioral sciences, such as sociology (e.g., Kirkegaard, 2016), education (e.g., Beavers et al., 2013), organizational research (e.g., Conway et al., 2003), and communication science (e.g., Park et al., 2002).

Precursors to contemporary EFA have been available almost as soon as correlation matrices could be calculated (see Mulaik, 2010 for a review). As Sir Francis Galton and Karl Pearson worked on mathematical models of correlation, that would ultimately yield the still-popular Pearson product moment correlation coefficient (Pearson, 1909), other scholars used these methods to calculate intercorrelation matrices to evaluate higher-order indices for interrelated variables. Charles Spearman, for example, built a higher-order model of intelligence (Spearman, 1904). Spearman's work was actually more reminiscent of the bifactor model approach, which has become quite popular recently, than EFA (see e.g., Sellbom & Tellegen, 2019). It was subsequent scholars who ultimately advocated for the EFA principles and methods that are frequently used today (e.g., Cattell, 1943; Thurstone, 1938).

This chapter emphasizes EFA. There are several forms of data reduction techniques, such as principal components analysis (PCA) and image factor analysis, which make different assumptions about the variances in the variables (i.e., indicators) being analyzed. We also focus specifically on latent variable models and do not provide coverage of other related methods of evaluating structure of variables, such as network analysis (e.g., exploratory graph analysis) or person-centered cluster analytical approaches. Furthermore, we do not cover confirmatory factor analysis (CFA) – a special case of structural equation modeling (SEM, see Chapter 25 in this volume) – though many issues pertaining to indicator selection and estimators apply to CFA as well. The "SEM Steps and Reporting Standards" in Chapter 25 of this volume apply to specifying, estimating, and evaluating a CFA model; Figure 25.2 and its associated narrative discussion provides a good example in that chapter. We further note that this chapter is mostly applied in nature, meaning that we do not take a mathematical approach to explaining the conceptual and practical foundations of EFA. Indeed, any reader interested in such foundations is referred to Mulaik's (2010) excellent book on this topic.

Definitions and Contrasting from Other Methods

Readers may rightfully wonder how EFA is different from some of these other data reduction methods just mentioned. We will contrast EFA from the two most common other alternatives (PCA and CFA). First, PCA and EFA are often confused as they are both similar forms of data reduction, and similar steps are applied in selecting the optimal structure – rotating solutions to simple structure, theoretical evaluation of competing structures, etc. However, it is important to note that, even if PCA and EFA methods often yield similar solutions, they are based on different assumptions about the underlying variances in the variables being analyzed.

EFA is based on the *common factor model* (e.g., Thurstone, 1947), which assumes each variable in a set of observed or measured variables (i.e., *indicators*) is a linear function of one or more unobserved (i.e., *latent*) factors as well as a residual factor unique to each variable (i.e., a *unique variance* component). Each latent variable is estimated through the variance common across the set of indicators (hence, common factors) and, specifically, that is being predicted by the latent variable. In other words, the underlying reason (or "cause") for a particular value on any observed indicator is the level of the underlying latent construct. The common factor model also considers residual factors that represent the unique variance of each indicator when the common variance has been accounted for; this unique variance is a combination of both systematic (or reliable) influences that are unrelated to the latent variable(s) as well as unsystematic (or unreliable) variances.

PCA, on the other hand, is not based on the common factor model as it does not parcel out shared and unique variances (Mulaik, 2010). Rather, PCA is a more simplistic procedure that attempts to maximize the amount of variance for which can be accounted in the indicators rather than making assumptions about causation. Brown (2014) points out that some scholars nonetheless argue that PCA might be advantageous to EFA because it is more simplistic mathematically, is less prone to problematic solutions, is not hampered by factor indeterminacy (i.e., component scores can be calculated more easily than factor scores), and PCA and EFA often yield similar results. However, as also noted by Brown (2014), other scholars (e.g., Fabrigar et al., 1999; Floyd & Widaman, 1995; see also Schmitt, 2011) have generally refuted these arguments because solutions are indeed dissimilar under various conditions (e.g., few indicators per factor and small communalities – the amount of variance accounted for in an indicator by all factors); more generally, analyses should be applied based on the underlying theoretical assumptions made about associations among variables. Moreover, because both EFA and CFA are based on the common factor model, EFA results are more likely to be supported by subsequent CFA in other samples (Floyd & Widaman, 1995; Schmitt, 2011).

The primary difference between EFA and CFA are in the names. Exploratory methods make no a priori assumptions about structure and are best suited for contexts in which the underlying structure of a set of variables is unknown. Confirmatory analyses, on the other hand, are explicitly testing one or several competing theoretical structures that are indeed known. Both are based on the common factor model, but one important difference in CFA is the reliance on the independent clusters model. In other words, a standard CFA model typically assumes one cause (i.e., latent factor) per indicator. EFA, on the other hand, makes no such assumption and estimates all latent factors as predictors for all indicators in the model; instead, it uses rotation methods (defined in a later section) to view a particular solution from a simple structure perspective. Finally, unlike EFA, the global evaluation of CFA models is largely based on the degree to which the specified model is consistent with the observed data (i.e., model fit) and statistical comparison to other theoretically plausible models. See Chapter 25 for detailed coverage of these issues in the broader SEM context.

General Applications

EFA can be useful in any context in which higher-order explanations of intercorrelations among a set of variables can be beneficial. Indeed, it has been used to articulate the structure of major psychological constructs, including intelligence (e.g., Thurstone & Thurstone, 1941) and personality (e.g., Cattell, 1945). A particularly common application of EFA is the examination of the internal structure of psychological test items (Brown, 2014), especially when no clear a priori theoretical structure exists. For instance, imagine a researcher has developed a new self-report questionnaire for assessing educational learning strategy – a multidimensional construct composed of multiple abilities (e.g., Berger & Karabenick, 2016). Therefore, the researcher, who has developed 25 test items to measure important features of learning strategies, conducts an EFA to determine the underlying structure of the test.

As another recently published example, Jokiniemi et al. (2021) developed a clinical nurse specialist core competency scale in a sample of nurses from a variety of Nordic countries. Because the underlying structure of the scale was unknown, they subjected 50 items to an EFA and decided on a four-factor structure. The respective loadings of items on the factors indicated four competency spheres of patient, nursing, organization, and scholarship, which formed a final scale.

Steps in Conducting EFA

In this section, we articulate the various steps in the applied use of EFA. For each, we discuss important issues that EFA users should consider and the general empirical literature to guide decisions. We also provide exemplary *R* code for practitioners to illustrate how to conduct the steps 2–5. A more detailed data example can be found in our Open Science Framework (OSF) repository (https://osf.io/srv8e/). The data were provided by Schödel et al. (2018) and consist of 312 observations of 60 extraversion items (four-point Likert scale) from the Big Five Structure Inventory (BFSI; Arendasy 2009); the BFSI measures the five-factor model of personality – a popular personality perspective in psychology. Extraversion is conceptualized as a broad individual differences trait domain with multiple specific trait facets (e.g., gregariousness, warmth, assertiveness), and thus, an extraversion item pool should be multifactorial. Throughout this chapter, we use these data to walk the reader through the basic steps of EFA; for a more detailed depiction, we encourage the reader to study the supplemental material in our OSF repository.

Step 1: Variable Selection

Every EFA begins with variable (indicator) selection, which is dictated by the purpose of the analysis. It is important to keep in mind that the results of EFA are completely bound by the variables included; there is no magic that will reveal some broad truth. Thus, EFA should not be used to articulate theory but, rather, thinking carefully about variable selection should precede the analysis. Of course, for certain applications, the researchers are bound to a particular variable pool. For instance, the evaluation of the internal structure of a psychological test is directly linked to the available items on that test.

It is very important that EFA users pay close attention to the nature of their indicators (e.g., scaling, distribution, degree of unidimensionality), as these properties have important implications for the selection of factor-extraction methods discussed later. Indeed, considerations for scaling/distribution of indicators will be covered under Step 3. In this first step, we consider some issues concerning the nature of indicators to which we believe EFA users should pay particular attention.

In an excellent article on factor analysis, Schmitt (2011) argues that indicators that are highly skewed can, for that reason alone, be highly correlated and indicative of an artefactual factor (see also Sellbom & Tellegen, 2019). This can be illustrated through a simple example. Let's consider two scales from the Minnesota Multiphasic Personality Inventory-2 – Restructured Form (MMPI-2-RF; Ben-Porath & Tellegen, 2008). Substance Abuse (SUB; seven items) and Anxiety (AXY; five items) measure two theoretically distinct constructs. We subjected its seven binary (true/false) items to an EFA using a robust weighted least-squares estimator in the modeling software Mplus 8.4 across two separate samples. The first sample consisted of 895 individuals from a community mental health center (Graham et al., 1999). The median endorsement frequency for these items was 27.5% (range: 9.3% to 51.6%). An EFA supported two factors (based on parallel analysis; see Step 4 later), with all seven SUB items loading on the first factor (median = 0.70; range: 0.45 to 0.87) and all AXY items loading on a second factor (median = 0.69; range: 0.48 to 0.79).

The second sample consisted of 336 individuals who had been administered the MMPI-2-RF as part of a pre-employment evaluation for a law enforcement position (Detrick et al., 2016) – a context in which endorsing substance abuse and anxiety symptoms is unlikely to occur due to either good psychological adjustment in such individuals or significant under-reporting. Indeed, the median endorsement frequency of these 12 items was 0.8% (range: 0.3–18.1%). The EFA suggested a clear one-factor solution with all but one of the items loading meaningfully on this factor (median = 0.84, range: 0.37–0.95); the only item that failed to reach a meaningful loading (0.28) was associated with the highest response endorsement (18.1%), and the item with the lowest meaningful loading (0.37) was associated with the second highest response rate (14.8%). Forcing a two-factor solution would result in an improper solution, with the single item left out of the one-factor model forming its own factor with a loading of 1.06. Thus, this example clearly demonstrates the effect that similar and extreme item skew can lead to theoretically inconsistent and artefactual factor solutions.

Indicator parceling is another important issue that EFA users should consider. Parceling refers to adding multiple indicators into a smaller set of aggregates to reduce model complexity. There is debate in the field about whether parceling is appropriate in factor analysis (e.g., Bandalos, 2008; Little et al., 2013; Marsh et al., 2013). Proponents for parceling argue for parcels being more reliable and distributionally sound indicators than the original ones, as well as the benefits of reducing model complexity and increase in statistical power (Little et al., 2013). Opponents, however, argue that parceling can mask potential problems associated with individual indicators, as poor indicator performance could be indicative of problematic content contributing construct-irrelevant variance (e.g., Bandalos, 2008; Marsh et al., 2013). We do not take a strong stance other than to say that if the indicators subjected to EFA meet the general goal of the analysis, and they are sufficiently unidimensional/reliable, indicator parceling is generally appropriate.

Step 2: Compute the Variance–Covariance Matrix

A factor analysis is a test of a variance–covariance matrix (or, in standardized terms, a correlation matrix). Every statistical software will calculate a variance–covariance matrix and subject this matrix to an EFA. In most cases, EFA users do not need to do this themselves. The statistical program will automatically do the conversion based on the instructions received. However, it is also possible for the applied user to calculate a variance–covariance matrix and directly subject this matrix to EFA to make adjustments to the correlations in a manner not possible when using raw indicator data (e.g., dis-attenuating correlations for range restriction or converting a correlation matrix to fit distributional assumptions); a full discussion of these issues is beyond the scope of our chapter. Here, we focus on some important assumptions about the variance–covariance matrix for EFA.

Prior to an EFA being conducted, the user should check whether the variance–covariance matrix meets the assumptions necessary to be subjected to the analysis. There are two common tests. First, the Kaiser–Meyer–Olkin (KMO) index of sampling adequacy directly allows for the examination of whether the variables to be included in the EFA are appropriate for factor analysis as a set. If the common variance across the indicators is too small, it is not meaningful to conduct an EFA. Hence, the KMO – a measure of the proportion of common variance across all indicators – states how well the set of indicators is suited for EFA. Values that are close to 1.0 are preferable (see Kaiser & Rice, 1974 for more guidance on KMO interpretation). The R package psych provides a function to calculate the KMO measure: psych::KMO(efa_data). For our example data on 60 extraversion items, the KMO measure is 0.92 and suggests that the data can be subjected to an EFA.

Another method to determine the suitability of an indicator set for EFA is the Bartlett's test of sphericity that can be used to ensure that the correlation matrix is not an identity matrix. Specifically, if the indicators are not related to one another, no informative structure can be detected. A chi-square test is calculated to test the null hypothesis that the indicators are orthogonal; a significant test, therefore, is evidence that they are not and that the underlying matrix has sufficient covariation to be suitable for EFA. In R, users can apply the cortest.bartlett function of the psych package – psych::cortest.bartlett(cor(efa_data), n = nrow(efa_data)), with cor(efa_data) calculating the correlation matrix and nrow(efa_data) returning the sample size. Bartlett's test of sphericity rejects the null hypothesis in our data example, so we deem our data suitable for an EFA.

Step 3: Factor Extraction

When conducting an EFA, researchers can choose between several factor-extraction methods (i.e., estimation methods). This decision can have substantial influence on the results – especially on the estimated factor loadings (Beauducel, 2001; De Winter & Dodou, 2012). Although it has often been argued that PCA relies on different model assumptions (see earlier discussion on this topic) and, therefore, should not be treated as an alternative extraction method (e.g., Farbrigar et al., 1999) as it may yield biased factor loadings (Widaman, 1993), it is not always clear which other method is preferable (based on the common factor model).

Principal Axis Factoring

Historically, principal axis factoring (PAF) has been the most popular extraction method and remains the preferred method for many researchers who use EFA (Conway & Huffcutt, 2003; Goretzko et al., 2021; Henson & Roberts, 2006; Howard, 2016). Its popularity may be explained by its conceptual resemblance to PCA and its importance in the early days of EFA usage (e.g., Holzinger, 1946). The basic idea of PAF is to adjust the PCA approach to the common factor model to account for measurement error and to consider unique variance components. Instead of decomposing the correlation (or variance–covariance) matrix to find principal components, it works with a so-called "reduced" correlation matrix that contains communality estimates on the diagonal. Hence, the principal axes (factors) obtained from this procedure are not able to explain all the variance of the manifest variables but only shared variances according to the common factor model (i.e., the variance components that are explained by the underlying latent factors that represent, for example, psychological constructs).

The initial estimation of the communalities used to generate the reduced correlation matrix is usually based on the squared multiple correlations (SMCs) among the indicators – the default in statistic programs (e.g., SPSS or R when using the psych library: psych::fa(efa_data, nfactors = 6, fm = "pa", SMC = TRUE)). After initially "guessing" the communalities, an eigenvalue decomposition is performed on the reduced correlation matrix (similar to PCA), and the resulting factor pattern can then be used to re-estimate the communalities. This iterative procedure is continued until a convergence criterion is fulfilled.

(Weighted) Least-Squares Approaches

With this procedure, PAF implicitly aims at finding the factor loadings that minimize the squared deviation between the diagonals of the reduced correlation matrix (which contains the communality estimates) and the reproduced correlation matrix (which consists of the model-implied correlations calculated from the estimated factor loadings; see Jöreskog et al., 2016 for more details). While PAF uses the iterative procedure to estimate loadings and communalities via eigenvalue decomposition, Harman and Jones's (1966) minimizing residuals factor analysis (Minres) determines the factor loadings by reproducing the off-diagonal elements of the correlation matrix as closely as possible (thus circumventing the problem of unique variance elements). The authors demonstrate that Minres and PAF result in the same factor solution if the communalities estimated by Minres are used for PAF (Harman & Jones, 1966).

As Minres minimizes the squared distance between the off-diagonal elements of the correlation matrix and the respective correlations implied by the factor model (i.e., the squared residuals), it provides the solution to the unweighted least-squares-fit function (Jöreskog et al., 2016). Accordingly, Minres yields equivalent results as an ordinary least-squares method and can be seen as a representative of diverse least-squares approaches, which all estimate the model parameters by optimizing a fitting function that compares the actual correlation matrix with a model-implied matrix. The different estimation methods – namely generalized least-squares, unweighted least-squares, and maximum-likelihood (ML) methods – can be formulated as weighted least-squares (WLS) approaches (Browne, 1977) and enable the researchers to test the goodness of fit by calculating common model fit indices (e.g., the RMSEA). Robust WLS or diagonally WLS methods have been developed in the context of SEM research to address the problem of biased standard error estimation in WLS approaches (for an overview, see DiStefano & Morgan, 2014). These approaches, including, for example, WLSMV (weighted least-squares mean and variance adjusted) methods, can also be used in EFA. However, robust WLS is often not selected as an extraction method, probably because EFA is usually used for exploring the data and not for model testing (compared to CFA) and, hence, proper standard error estimation is rarely considered by its users.

One of the advantages of these least-squares approaches is that they do not carry distributional assumptions about the indicators and are, therefore, applicable to all types of variables. As described later, research has indicated that these approaches can be particularly useful for ordered categorical or binary variables (e.g., Rhemtulla et al., 2012). However, a WLS analysis usually needs (slightly) larger sample sizes (e.g., Li, 2016; Rhemtulla et al., 2012) and is not the preferred method when using normally distributed indicators.

Maximum-Likelihood Estimation

While general least-squares approaches come without distributional assumptions, ML estimation puts a stronger focus on the data-generating process and is, therefore, arguably a more sophisticated approach to factor analysis – especially since it treats the unique variances as formal model parameters that have to be also estimated (Everitt & Hothorn, 2011). Usually, multivariate normality is assumed but, theoretically, other distributional assumptions can also be made for ML estimation (e.g., Wedel & Kamakura, 2001).

For ML estimation, a fitting function that is closely related to the likelihood function is minimized with respect to the loading parameters as well as the unique variances. The likelihood function indicates how plausible specific parameter values are given the observed data; in this case, that means the plausibility of specific factor loadings and unique variances. Accordingly, the aim of this estimator is to maximize the likelihood that the final set of parameters map onto the observed data (hence, "maximum likelihood"). Moreover, even though the ML estimation of parameters is fairly robust against violations of the normality assumption (e.g., Jöreskog et al., 2016), standard errors and respective significance tests can deteriorate when the actual data-generating process differs from the assumed one. Therefore, several adjustments for robust ML estimation have been developed (e.g., Yuan & Bentler, 1998).

Comparison of Factor-Extraction Methods

Selecting one of these estimation methods for an EFA can be challenging as their precision and stability vary across different data conditions. PAF is sometimes favored as it produces fewer Heywood cases (i.e., cases in which unique variances are estimated to be negative or correlations estimated to be greater than one) compared to ML estimation (De Winter & Dodou, 2012). PAF also does not require multivariate normality underlying the indicators. However, the initial communality estimates can heavily influence the outcome of PAF; using the complete variances as communality estimates often yields inflated parameter estimates, while using the SMC approach may cause negative eigenvalues (Gorsuch, 1983). Furthermore, PAF does not allow for a direct replication with CFA (for which the default estimator is typically ML) and does not provide fit indices to evaluate model fit. Hence, several authors advocate not to rely on PAF (Conway & Huffcutt 2003; Fabrigar et al., 1999; Goretzko et al., 2021) but rather to use ML estimation especially when multivariate normality can be assumed. Because EFA results should typically be replicated and validated using CFA on a new sample, a likelihood-based estimation procedure seems to be the most suitable.

When the multivariate normality assumption is violated (e.g., when data are based on indicators with few categories), WLS parameter estimation based on polychoric correlations may be an appropriate alternative to ML estimation (Barendse et al., 2015; Schmitt, 2011); it also can be used in CFA and, hence, for direct cross-validation. EFA users should examine their data carefully and evaluate whether a normality assumption holds. Regardless, when ordinal indicators with fewer than five categories are used, WLS is preferred over ML estimation – particularly robust weighted least squares or unweighted least squares (Beauducel & Herzberg, 2006; Goretzko et al., 2021; Li, 2016; Rhemtulla et al., 2012). The "fa" function of the psych package offers numerous estimation methods. Users can select the preferred method by setting the argument "fm", for example, when performing WLS estimation: psych::fa(efa_data, nfactors = 6, fm = "wls"). Since our example data set consists of four-point Likert items that have to be considered as ordinal variables (see also Beauducel & Herzberg, 2006), we decided to rely on WLS estimation (statistical tests and graphical inspection also suggest that multivariate normality is questionable for our data, see https://osf.io/srv8e/).

Step 4: Factor Retention

Selecting the optimal number of factors to retain constitutes a key decision in EFA. Before estimating the loadings and unique variances, the researcher needs to determine the dimensionality (or number of latent factors). Although theoretical considerations should also be taken into account in this decision-making process, this number is primarily inferred from objective data. Over the years, several factor-retention criteria have been developed to estimate the number of latent factors underlying the correlation matrix of the manifest indicators.

Eigenvalues: Kaiser–Guttman and Empirical Kaiser Criterion

Eigenvalues are central characteristics of a matrix that, in the case of correlation matrices, indicate how much variance in the manifest variables can be explained by the respective eigenvectors (i.e., the principal components in PCA). Therefore, eigenvalues of the correlation matrix (or the reduced correlation matrix in PAF) play a central role in determining the number of factors to retain in EFA. The well-known Kaiser–Guttman rule (Kaiser, 1960), often referred to as eigenvalue-greater-one-rule, suggests retaining as many factors as there are eigenvalues greater than one. At the population level, the correlation matrix under the null model (no underlying factors) is simply an identity matrix and all eigenvalues are one. Accordingly, the rationale of the Kaiser–Guttman rule is that an underlying factor should explain more variance than a single variable and should have a corresponding eigenvalue greater than one.

As Breaken and van Assen (2017) explain, this idea may be reasonable on a population level, but is flawed on a sample level due to sampling error. Therefore, the authors developed a new version of this rule – the empirical Kaiser criterion (EKC) – that takes into account the sample size as well as the size of previous eigenvalues when calculating reference eigenvalues that are compared with the empirical eigenvalues (e.g., the second reference eigenvalue is adjusted to account for a very large first eigenvalue corresponding to a dominant first factor). In other words, EKC provides different reference values for each observed eigenvalue (instead of comparing all eigenvalues with the fixed value of one) and promises to be less prone to sampling error. It suggests retaining factors whose eigenvalues are greater than the calculated reference eigenvalues, considering the sample size, the number of indicators, and all previous eigenvalues. Braeken and van Assen (2017) differentiate the restricted EKC (where the reference eigenvalues are at least one) and an unrestricted version with reference eigenvalues that can be even smaller than one.

Scree Test

Another popular method of determining the number of factors is the scree test (Cattell, 1966) and it is also based on the empirical eigenvalues. The idea behind this method is to plot the eigenvalues in a descending order and to determine an "elbow" in this plot, where the change from one eigenvalue to the subsequent eigenvalue is considerably smaller than the difference between the two prior eigenvalues. The assumption is that all factors corresponding to the eigenvalues before this "elbow" can explain substantial amounts of variance while all factors from this position and onwards are insufficient for this purpose. Ultimately, the visual inspection and interpretation of this scree plot is quite subjective.

Parallel Analysis and Comparison Data

The improvement in computational resources have fostered the applicability of simulation-based factor-retention approaches. Parallel analysis (PA; first implemented by

Horn, 1965) is the best-known factor-retention approach that uses simulated data for comparison. The basic premise of PA is to generate reference values for the empirical eigenvalues based on several simulated data sets of the same size and number of indicators as the empirical data set. After simulating B data sets based on the null model (i.e., no underlying latent factors), the mean of the B first eigenvalues is compared to the first empirical eigenvalue, the mean of the B second eigenvalues is compared to the second empirical eigenvalue, and so on. PA suggests retaining factors as long as the empirical eigenvalue is greater than the reference eigenvalue. Instead of using the mean to aggregate the respective eigenvalues of the B data sets, arbitrary percentiles of the eigenvalue distribution can be taken as the reference value (often the 95% percentile). There are also implementations of PA that are based on the eigenvalues of the reduced correlation matrix (see also the comparison of PCA and EFA) and PA varieties using bootstrapped instead of simulated data. Lim and Jahng (2019) provide a more detailed overview of the different versions of PA and their performance under various data conditions.

Ruscio and Roche (2012) developed the comparison data (CD) approach that combines the simulation of comparison data sets (similar to PA) with the model-testing perspective of CFA (see also the section on model fit indices below). Contrary to PA, the simulated data sets do not represent a null model but are based on different factor models while also reflecting the marginal distributions of the indicators. For each factor solution and each comparison data set, the root-mean-squared error (RMSE) between the empirical eigenvalues and the respective comparison eigenvalues is calculated. That is, if B comparative data sets are simulated per factor solution, B RMSE values per number of factors are obtained. The CD method then tests whether the RMSE values of a two-factor solution are, on average, significantly smaller than those of a one-factor solution. Mann–Whitney U tests are conducted with subsequent numbers until no "significant" improvement is indicated. To avoid underfactoring, the authors suggest an alpha level of 0.30 as a threshold for significance.

Minimum Average Partial Test

Velicer (1976) developed the minimum average partial (MAP) test that aims at determining the number of components to retain in PCA based on averaged, squared partial correlations of the indicators. Although it was designed for PCA, the MAP test is frequently used in the context of EFA (Goretzko et al., 2021). The basic premise is to determine the number of components for which the squared correlations of the indicators are minimal, on average, after the common variance explained by the principal components is controlled for ("partialed out").

Hull Method

The hull method by Lorenzo-Seva et al. (2011) consists of three major steps. First, a set of factor solutions is selected for which a model fit index is calculated (the authors suggest using the comparative fit index). Then, the fit index is plotted against the corresponding degrees of freedom for each factor solution. Subsequently, an elbow in the upper boundary of the convex hull of the plotted points is detected to determine at which point increasing the number of factors does not substantially improve upon the model fit. Unlike the scree test, the position of the elbow in the upper hull can be calculated, making this a less subjective approach. Due to its model-comparison perspective, the hull method necessitates the use of ML or least squares estimators to calculate model fit indices.

Sequential Chi-Square Tests

When using ML EFA, it is possible to test whether a specific number of factors *k* is sufficient to explain the common variance of an indicator set. If that is the case, a test statistic proportional to the ML fitting function is approximately chi-square distributed (e.g., Everitt & Hothorn, 2011) and the null hypothesis that *k* factors are sufficient can be tested. This procedure is repeated with subsequent numbers of factors ($k = 1, 2, 3, \ldots$) until the null hypothesis holds.

Fit Indices and Information Criteria

There are also authors who view factor retention as a model selection problem (e.g., Preacher et al., 2013) and, therefore, rely on relative and absolute measures of model fit. When likelihood-based EFA is conducted, information criteria, such as the Akaike information criterion (Akaike, 1987) or the Bayesian information criterion (Schwarz, 1978), can be used to determine which number of factors represents the empirical relations more accurately. As an alternative to information criteria, fit indices known from model testing in the context of CFA (or structural equation modeling; see Chapter 25 in this volume) can also be used to compare different factor solutions with each other (see Preacher et al., 2013 for more details). Some recent scholars have published simulation data that question the utility of model fit indices for the purposes of factor retention, however (Auerswald & Moshagen, 2019; Montoya & Edwards, 2021).

Factor Forest

Recently, a new simulation- and machine learning-based approach for factor retention has been developed by Goretzko and Bühner (2020). The basic idea of the factor forest is to simulate data under all important data conditions of an application context (i.e., considering common sample sizes, realistic ranges for the number of latent factors and the number of manifest indicators, common loading patterns and communalities, etc.) and then to extract specific data characteristics for each simulated data set (e.g., eigenvalues and matrix norms of the correlation matrix). These data characteristics, and the known number of latent factors (the true dimensionality is known since the data are simulated), are then treated as input (independent variables) and target variables (dependent variable or criterion) of a machine learning model that "learns" how the data characteristics and the number of factors are interlinked.

The trained model is then able to predict the number of factors given the observed data characteristics of an empirical data set. As this procedure is computationally very costly, Goretzko and Bühner (2020) provide a pre-trained model that was trained on nearly 500,000 data sets based on multivariate normality and between one and eight latent factors; the trained model and the analysis scripts can be retrieved from an OSF repository – https://osf.io/mvrau/ or from our repository with a simplified R script – https://osf.io/srv8e/).

Comparison of Factor-Retention Criteria

Although the Kaiser–Guttman rule, the scree test, and PA are the most popular methods to determine the number of factors (Goretzko et al., 2021), simulation studies suggest that only the latter provides comparably good estimates, and the other methods are often not able to retain the correct number of factors (Auerswald & Moshagen, 2019; Fabrigar et al., 1999; Goretzko et al., 2021; Schmitt et al., 2018). Therefore, PA is seen as the “gold standard” of factor retention (e.g., Braeken & van Assen, 2017; Schmitt et al., 2018); this may also be explained by its relative robustness against distributional assumptions (Dinno, 2009). However, some modern alternatives (e.g., the CD method, EKC, or hull method) have shown advantages over PA in some data conditions (Braeken & van Assen, 2017; Lorenzo-Seva et al., 2011; Ruscio & Roche, 2012). This is why several authors agree on consulting more than one factor-retention criterion (Fabrigar et al., 1999; Goretzko et al., 2021) or using combination rules (e.g., Auerswald & Moshagen, 2019). The pre-trained factor forest model showed very high accuracy in Goretzko and Bühner’s (2020) study and may be a more convenient alternative for practitioners as it internally weighs different methods (PA, EKC, CD). As mentioned earlier, the use of model fit indices in factor retention is less defensible (Auerswald & Moshagen, 2019; Montoya & Edwards, 2021). Ultimately, searching for a perfect factor solution in a myriad of available criteria may sometimes be an exercise in futility (e.g., Cattell, 1966). EFA users might, therefore, also consider (in addition to the aforementioned objective recommendations) theoretical utility and a strive towards parsimony in this venture (e.g., Schmitt et al., 2018).

The R packages psych and EFAtools provide functions for the most common criteria (e.g., PA: psych::fa.parallel(efa_data, fm = “wls”) or CD: EFAtools::CD (efa_data, n_factors_max = 8)). For our data example, we compared several factor-retention criteria (see https://osf.io/srv8e/ for the full R code). The most reliable methods (see, Auerswald & Moshagen, 2019; Goretzko and Bühner, 2020) – PA, EKC, CD, MAP test, and the factor forest – suggested between five and six factors. Since theoretical considerations (the BFSI claims to measure six facets of the extraversion trait domain with ten items each) speak in favor of a six-factor solution, we retained six latent variables.

Step 5: Factor Rotation

In EFA, all indicators are explained by the set of retained latent factors as dependent variables in a linear regression system. The factor loadings (i.e., the regression parameters) are standardized and expressed in a correlation metric. However, in the initial, unrotated solution, the matrix containing these loadings (factor–indicator correlations) often does not adhere to a clean pattern and makes the interpretation of the factor solution quite difficult. For this reason, factor-rotation methods have been developed to elucidate a more interpretable solution, the so-called *simple structure* (i.e., each indicator loads high on its associated factor and low on all other factors – ideally all cross-loadings are zero) – an idea that was originally presented by Thurstone (1947). More specifically, when estimating the factor loadings and unique variances, the problem of rotation indeterminacy arises (Mulaik, 2010). That is, the loading pattern or loading matrix is only determined up to an arbitrary rotation, and, therefore, selecting an appropriate rotation method solely depends on theoretical considerations and the interpretability of the resulting factor solution. In other words, there is no data-driven way to decide how to rotate the factor solution (see also Browne, 2001; Goretzko et al., 2021).

Orthogonal vs. Oblique Rotation

To obtain such an interpretable solution, two different types of rotations can be used – orthogonal and oblique rotation techniques. Historically, orthogonal rotation methods, which yield uncorrelated factor solutions (all between-factor correlations are constrained to be zero), have been applied more frequently. An advantage of orthogonal or uncorrelated factors is that the respective constructs are clearly distinguishable and that relations between them and third variables can be evaluated independently from each other. However, this process might, in many instances, distort the natural structure of the data when constructs are indeed correlated. Accordingly, oblique rotations, allowing factors to correlate, may be a more appropriate assumption for most social and behavioral research phenomena.

Varimax

The most popular orthogonal rotation method is called varimax (Kaiser, 1958). As the name suggests, the varimax criterion rotates the initial factor solution in a way that maximizes the variance of the squared loadings by columns (i.e., the variance of the squared loadings is maximized for each factor). Hence, this rotation yields rather extreme loadings (either high loadings or very small loadings on each factor).

Quartimax and Equamax

Quartimax is another member of the orthomax family of criteria (Harman, 1976), which includes several orthogonal rotation techniques (e.g., inter alia varimax). Contrary to varimax, it focuses on the row-wise complexity and favors patterns for which each variable has as many zero-loadings as possible. This process leads to an insensitivity to a strong first factor; this is why quartimax often yields a general factor (or a strong first factor and smaller or more trivial second and third factors, etc.). Equamax (see, for example, Kaiser, 1974) is a combination of varimax and quartimax criteria that tries to minimize the number of large loadings per factor and the number of large loadings per variable at the same time.

Promax

One of the most prominent oblique rotation methods is promax (Hendrickson & White, 1964) – a two-stage method that first applies an orthogonal rotation (e.g., varimax) and then subsequently performs the actual oblique rotation. In this process, larger loadings are enhanced compared to smaller loadings by matching the factor loading pattern as closely as possible to an exponentiated version of itself. Usually, the orthogonal loadings are raised to the power of four (e.g., the default setting in the psych package in R; see also the discussion of Hendrickson & White, 1964 on why four was chosen as the default setting for promax), but the exponent can be changed depending on theoretical considerations. It is important to note that a larger exponent will result in larger between-factor correlations.

Oblimin (Family)

Another oblique rotation method that is frequently used in psychological research is called oblimin (Clarkson & Jennrich, 1988). Strictly speaking, it is a family of oblimin methods that includes different rotation techniques, such as quartimin (the oblique generalization of quartimax) or covarimin (the oblique generalization of varimax) as special cases (Clarkson & Jennrich, 1988). A parameter (often named δ [in SPSS] or γ [in R]), which controls the "obliqueness" of the rotated factor solution, determines which rotation of the oblimin family is applied. Jennrich (1979) demonstrated that positive parameter values can be inadmissible when performing oblique rotation; this is why the default value in statistical programs like SPSS and R (we refer to the psych package and the GPArotation package) is zero and corresponds to the quartimin criterion. Quartimin rotation yields a more oblique solution as it minimizes the row-wise complexity by introducing higher inter-factor correlations; decreasing the parameter value (selecting a more negative value) yields a less oblique or more orthogonal solution. Oblimin is selected as the default rotation method in psych.

Geomin

Geomin (Yates, 1987) is a newer oblique rotation technique (there is an orthogonal version as well, e.g., Browne, 2001) that minimizes an objective function based on row-wise geometric means of the squared factor loadings. Thus, geomin focuses on row-wise complexity (i.e., it tries to minimize the number of factors that are needed to explain the variance of each indicator variable). Geomin is the default rotation in Mplus and shows comparably good results when little is known about the true loading pattern (Asparouhov & Muthén, 2009).

Crawford–Ferguson Family

Crawford and Ferguson (1970) presented a general objective function or rotation criterion that is a weighted sum of row and column complexity. Several well-known rotation techniques can be integrated in their general framework. In fact, the

Crawford–Ferguson (CF) family is equivalent to the orthomax family in the orthogonal case (Crawford & Ferguson, 1970) but yield different results when oblique rotations are considered (Browne, 2001). As a counterpart of quartimax (focus on row-wise complexity), Crawford and Ferguson (1970) introduced CF–facparsim that aims at factor parsimony (focus on column-wise complexity) and, therefore, tries to minimize the number of variables that load on each factor.

Comparison of Rotation Methods

In current research, many EFA users rely on varimax criteron for orthogonal rotation as well as on promax and oblimin/quartimin criteria for oblique rotation (Fabrigar et al., 1999; Goretzko et al., 2021). As pointed out earlier, there is no "correct" way of rotating the initial factor solution. However, many researchers recommend oblique rotation techniques since ruling out between-factor correlations in advance seems to be less plausible in social and behavioral science research (Conway & Huffcutt 2003; Fabrigar et al. 1999; Goretzko et al., 2021).

There are very few recommendations when it comes to choosing an oblique rotation method, though. Simulation studies (e.g., Sass and Schmitt, 2010) suggest that researchers should rely on CF–equamax or CF–facparsim when they expect factor loading patterns with high complexity (i.e., several substantial cross-loadings), whereas geomin or CF–quartimin seem to be more appropriate when patterns closer to simple structure can be assumed. Browne (2001) advocates for trying out different rotation methods (ideally on different subsamples if the data set is large enough for splitting) and to compare the results with regard to stability (if more than one subsample is used) and interpretability. He further suggests comparing a member of the CF family (e.g., CF–equamax) and geomin.

In modern software solutions, a variety of these rotation methods are implemented. The fa function of the psych package, for example, offers numerous options that users can select via the "rotate" argument – psych::fa(efa_data, nfactors = 6, fm = "wls", rotate = "Promax"). For our data example (https://osf.io/srv8e/), we also illustrated the two-step approach – first estimating all parameters for an unrotated solution and then applying a rotation method to increase the interpretability of the factor solution. Comparing the results of orthogonal varimax and oblique quartimin, inter-factor correlations seem to foster interpretability for our exemplary data; in other words, it seems to be reasonable to assume correlated facets, especially as all 60 items are considered to be indicators of the same personality trait (extraversion).

Step 6: Interpretation

The final step in the EFA process is to provide a theoretical interpretation of the solution. Because the analysis is, by definition, exploratory, the theoretical evaluation comes last and is needed to provide meaning to the resulting structure. There are three common vectors of information that EFA users consider in interpretation: factor loadings, communalities, and factor correlations.

Factor Loadings

The first consideration is the factor loadings; these represent the relationship between the latent factor and an indicator and, more specifically, the degree of variance that the factor accounts for in the indicator. Considering the initial, unrotated factor solution, squared factor loading represents the explained variance in the indicator. The pattern of factor loadings is used to provide meaning to the latent factors in EFA. The indicators with the largest and most distinct loadings are typically considered in the interpretation of the theoretical underpinnings of the latent variable.

There is no universally agreed upon threshold for what constitutes a sufficiently large factor loading for it to be considered meaningful. In psychology, and specifically evaluation of psychological tests at the item level, approximately 0.30–0.40 tends to be considered the lower bound for a meaningful loading; 0.50+ is considered large and substantial (e.g., Gorsuch, 1983). Furthermore, factors that are defined by many cross-loadings (or solely defined by them) are usually not meaningful and signal to the EFA user that too many factors have been extracted or some indicators are poor. A more extreme manifestation of this phenomenon is a “bloated specific” factor (Cattell & Tsujioka, 1964), with one or two very large loadings of indicators on one factor when a broader group of the similar variables are already represented in the remainder of the factor solution. Finally, indicators that have large loadings on more than one factor, unless theoretically indicated as representing variables with clear multiple causes identified in the factor solution (e.g., interstitial variables; see Krueger, 2013, for an example discussion in the personality literature), should also be candidates for elimination as poorly functioning variables (e.g., Brown, 2014).

Communalities

The second consideration on the overall evaluation of a factor solution is the communalities. A communality (h^2) is the total proportion of variance explained in an indicator by all retained factors, whereas $1 – h^2$ is the residual – the proportion of the systematic and unsystematic variance that is unique to the indicator. High communalities typically mean that the factor solution can account for most of the systematic variance in the indicators, whereas low communalities might reflect that the indicators are of lesser importance to the structure being evaluated – these should be considered for removal from the analysis (e.g., Brown, 2014) unless counter-indicated for theoretical reasons (e.g., reduces critical content coverage).

Factor Correlations

As for the final consideration, factor correlations indicate the degree of overlap between the latent factors that have been extracted in the EFA and rotated with an oblique method. These correlations should also be interpreted with theory in mind as there are no thresholds for what constitutes a meaningful correlation. If the emerging latent constructs in an EFA are conceptually expected to be relatively distinct (e.g., positive and negative emotions), smaller correlations are expected. On the other hand, if the constructs are conceptually expected to converge (e.g., impulsivity and risk taking), larger correlations are expected. Extremely high correlations (e.g., 0.80–0.90+) likely reflect significant redundancies in latent constructs and point towards a factor solution with a smaller number of factors.

In our data example, we found that individual items predominantly loaded meaningfully and relatively distinctly onto six latent factors reflecting warmth, gregariousness, assertiveness, drive, adventurousness, and cheerfulness. The latent factors were intercorrelated, as expected, but also distinct (all inter-factor correlations *r*s < 0.47 when applying quartimin rotation). Communality estimates also indicated that a meaningful proportion of variance was captured in each of the items, with only one exception. Overall, this structure was consistent with theoretical expectations associated with the extraversion trait domain.

Conclusion

This chapter has provided an introduction to the basics of applied EFA. Our goal was to review foundations for and steps associated with conducting an EFA in research. Specifically, we carefully considered each EFA step and described various considerations of which the EFA user should be mindful, including the fact that numerous options and, therefore, researcher degrees of freedom exist for each of these steps. Users should also be aware that many choices they have to make are not only of statistical nature but highly depend on the research questions being addressed, theoretical considerations in general, and the nature of the measured indicators. We believe that EFA is a powerful statistical method for the exploration of higher-order structure, but it is not straightforward and requires many decisions, with the incorrect ones possibly yielding biased results. We hope this guide will therefore be useful to the reader as they choose to apply this method in their research.

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*, 317–332.

Arendasy, M. (2009) *BFSI: Big-Five Struktur-Inventar (Test & Manual)*. Mödling, Schuhfried GmbH.

Asparouhov, T. & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling*, *16*(3), 397–438.

Auerswald, M. & Moshagen, M. (2019). How to determine the number of factors to retain in exploratory factor analysis: A comparison of extraction methods under realistic conditions. *Psychological Methods*, *24*(4), 468–491. https://doi.org/10.1037/met0000200

Bandalos, D. L. (2008). Is parceling really necessary? A comparison of results from item parceling and categorical variable methodology. *Structural Equation Modeling*, *15*(2), 211–240.

Barendse, M. T., Oort, F. J., & Timmerman, M. E. (2015). Using exploratory factor analysis to determine the dimensionality of discrete responses. *Structural Equation Modeling*, *22*(1), 87–101.

Beauducel, A. (2001). On the generalizability of factors: The influence of changing contexts of variables on different methods of factor extraction. *Methods of Psychological Research Online*, *6*(1), 69–96.

Beauducel, A. & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling*, 13(2), 186–203.

Ben-Porath, Y. S. & Tellegen, A. (2008). *Minnesota Multiphasic Personality Inventory-2 Restructured Form: Manual for Administration, Scoring and Interpretation*. University of Minnesota Press.

Berger, J. L. & Karabenick, S. A. (2016). Construct validity of self-reported metacognitive learning strategies. *Educational Assessment*, *21*(1), 19–33. https://doi.org/10.1080/10627197.2015.1127751

Braeken, J. & van Assen, M. A. (2017). An empirical Kaiser criterion. *Psychological Methods*, 22(3), 450–466. https://doi.org/10.1037/met0000074

Beavers, A. S., Lounsbury, J. W., Richards, J. K., et al. (2013). Practical considerations for using exploratory factor analysis in educational research. *Practical Assessment, Research, and Evaluation*, *18*(1). https://doi.org/10.7275/qv2q-rk76

Brown, T. A. (2014). *Confirmatory Factor Analysis for Applied Research*. Guilford Press.

Browne, M. W. (1977). Generalized least-squares estimators in the analysis of covariance structures. In D. J. Aigner & A. S. Goldberger (eds.), *Latent Variables in Socio-Economic Models* (pp. 205–226). North-Holland.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*(1), 111–150. https://doi.org/10.1207/S15327906MBR3601_05.

Cattell, R. B. (1943). The description of personality: Basic traits resolved into clusters. *Journal of Abnormal and Social Psychology*, *38*, 476–506.

Cattell, R. B. (1945). The description of personality: Principles and findings in a factor analysis. *The American Journal of Psychology*, *58*(1), 69–90.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10

Cattell, R. B. & Tsujioka, B. (1964). The importance of factor-trueness and validity, versus homogeneity and orthogonality, in test scales. *Educational and Psychological Measurement*, 24(1), 3–30.

Clarkson, D. B. & Jennrich, R. I. (1988). Quartic rotation criteria and algorithms. *Psychometrika*, *53*, 251–259.

Conway, J. M. & Huffcutt, A. I. (2003). A review and evaluation of exploratory factor analysis practices in organizational research. *Organizational Research Methods*, *6*(2), 147–168.

Crawford, C. B. & Ferguson, G. A. (1970). A general rotation criterion and its use inorthogonal rotation. *Psychometrika*, *35*, 321–332.

De Winter, J. C. & Dodou, D. (2012). Factor recovery by principal axis factoring and maximum likelihood factor analysis as a function of factor pattern and sample size. *Journal of Applied Statistics*, *39*(4), 695–710.

Detrick, P., Ben-Porath, Y. S., & Sellbom, M. (2016). Associations between MMPI-2-RF (restructured form) and Inwald Personality Inventory (IPI) scale scores in a law enforcement pre-employment screening sample. *Journal of Police and Criminal Psychology*, *31*, 81–95.

Dinno, A. (2009). Exploring the sensitivity of Horn's parallel analysis to the distributional form of random data. *Multivariate Behavioral Research*, *44*(3), 362–388. https://doi.org/10.1080/00273170902938969

DiStefano, C., & Morgan, G. B. (2014). A comparison of diagonal weighted least squares robust estimation techniques for ordinal data. *Structural Equation Modeling*, *21*(3), 425–438.

Everitt B. & Hothorn T. (2011) Exploratory factor analysis. In *An Introduction to Applied Multivariate Analysis with R* (pp. 135–161). Springer. https://doi.org/10.1007/978-1-4419-9650-3_5

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*(3), 272–299. https://doi.org/10.1037/1082-989X.4.3.272.

Floyd, F. J. & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, *7*(3), 286–299.

Goretzko, D. & Bühner, M. (2020). One model to rule them all? Using machine learning algorithms to determine the number of factors in exploratory factor analysis. *Psychological Methods*, *25*(6), 776–786. https://doi.org/10.1037/met0000262

Goretzko, D., Pham, T. T. H., & Bühner, M. (2021). Exploratory factor analysis: Current use, methodological developments and recommendations for good practice. *Current Psychology*, *40*(1), 3510–3521. https://doi.org/10.1007/s12144-019-00300-2

Gorsuch, R. L. (1983). *Factor Analysis*. Erlbaum.

Graham, J. R., Ben-Porath, Y. S., & McNulty, J. L. (1999). *MMPI-2 Correlates for Outpatient Mental Health*. University of Minnesota Press.

Harman, H. H. (1976). *Modern Factor Analysis*, 3rd ed. University of Chicago Press.

Harman, H. H. & Jones, W. H. (1966). Factor analysis by minimizing residuals (minres). *Psychometrika*, *31*, 351–368. https://doi.org/10.1007/BF02289468

Hendrickson, A. E. & White, P. O. (1964). Promax: A quick method for rotation to oblique simple structure. *British Journal of Statistical Psychology*, *17*(1), 65–70. https://doi.org/10.1111/j.2044-8317.1964.tb00244.x

Henson, R. K. & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement*, *66*(3), 393–416.

Holzinger, K. J. (1946). A comparison of the principal-axis and centroid factor. *Journal of Educational Psychology*, 37(8), 449–472. https://doi.org/10.1037/h0056539

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. https://doi.org/10.1007/BF02289447

Howard, M. C. (2016). A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve?. *International Journal of Human-Computer Interaction*, *32*(1), 51–62.

Jennrich, R. I. (1979). Admissible values of γ in direct oblimin rotation. *Psychometrika*, *44*, 173–177.

Jokiniemi, K., Pietilä, A. M., & Mikkonen, S. (2021). Construct validity of clinical nurse specialist core competency scale: An exploratory factor analysis. *Journal of Clinical Nursing*, *30*(13–14), 1863–1873.

Jöreskog K. G., Olsson U. H., Wallentin F. Y. (2016) Exploratory factor analysis (EFA). In *Multivariate Analysis with LISREL. Springer Series in Statistics* (pp. 257–282). Springer. https://doi.org/10.1007/978-3-319-33153-9_6

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*, 187–200.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*(1), 141–151. https://doi.org/10.1177/001316446002000116

Kaiser, H. F. (1974). A note on the equamax criterion. *Multivariate Behavioral Research*, *9*(4), 501–503.

Kaiser, H. F. & Rice, J. (1974). Little jiffy, mark IV. *Educational and Psychological Measurement*, *34*(1), 111–117.

Kirkegaard, E. O. (2016). Some new methods for exploratory factor analysis of socioeconomic data. *Open Quantitative Sociology & Political Science*, *1*(1), November 7. https://doi.org/10.26775/OQSPS.2016.11.07

Krueger, R. F. (2013). Personality disorders are the vanguard of the post-DSM-5.0 era. *Personality Disorders: Theory, Research, and Treatment*, *4*(4), 355–362. https://doi.org/10.1037/per0000028

Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936–949.

Lim, S. & Jahng, S. (2019). Determining the number of factors using parallel analysis and its recent variants. *Psychological Methods*, *24*(4), 452–467. https://doi.org/10.1037/met0000230

Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods*, *18*(3), 285–300. https://doi.org/10.1037/a0033266

Lorenzo-Seva, U., Timmerman, M. E., & Kiers, H. A. L. (2011). The hull method for selecting the number of common factors. *Multivariate Behavioral Research*, *46*(2), 340–364. https://doi.org/10.1080/00273171.2011.564527

Marsh, H. W., Lüdtke, O., Nagengast, B., Morin, A. J., & Von Davier, M. (2013). Why item parcels are (almost) never appropriate: Two wrongs do not make a right – Camouflaging misspecification with item parcels in CFA models. *Psychological Methods*, *18*(3), 257–284.

Montoya, A. K. & Edwards, M. C. (2021). The poor fit of model fit for selecting number of factors in exploratory factor analysis for scale evaluation. *Educational and Psychological Measurement*, 81(3), 413–440.

Mulaik, S. A. (2010). *Foundations of Factor Analysis*. CRC Press.

Park, H. S., Dailey, R., & Lemus, D. (2002). The use of exploratory factor analysis and principal components analysis in communication research. *Human Communication Research*, *28*(4), 562–577.

Pearson, K. (1909). Determination of the coefficient of correlation. *Science*, *30*(757), 23–25.

Preacher, K. J., Zhang, G., Kim, C., & Mels, G. (2013). Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective. *Multivariate Behavioral Research*, *48*(1), 28–56. doi:10.1080/00273171.2012.710386

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373. https://doi.org/10.1037/a0029315.

Ruscio, J. & Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. *Psychological Assessment*, *24*(2), 282–292. https://doi.org/10.1037/a0025697

Sass, D. A. & Schmitt, T. A. (2010). A comparative investigation of rotation criteria within exploratory factor analysis. *Multivariate Behavioral Research*, *45*(1), 73–103. https://doi.org/10.1080/00273170903504810.

Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment*, *29*(4), 304–321.

Schmitt, T. A., Sass, D. A., Chappelle, W., & Thompson, W. (2018). Selecting the "best" factor structure and moving measurement validation forward: An illustration. *Journal of Personality Assessment*, *100*(4), 345–362.

Schoedel, R., Au, J. Q., Völkel, S. T., et al. (2018) Digital footprints of sensation seeking. *Zeitschrift Für Psychologie*, *226*(4), 232–245.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Sellbom, M. & Tellegen, A. (2019). Factor analysis in psychological assessment research: Common pitfalls and recommendations. *Psychological Assessment*, *31*(12), 1428–1441. https://doi.org/10.1037/pas0000623

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

Thurstone, L. L. (1938). *Primary Mental Abilities*. University of Chicago Press.

Thurstone, L. L. (1947). *Multiple Factor Analysis*. University of Chicago Press.

Thurstone, L. L. & Thurstone, T. G. (1941). Factorial studies of intelligence. *Psychometric Monographs*, 2, 94.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, *41*(3), 321–327. https://doi.org/10.1007/BF02293557

Wedel, M. & Kamakura, W. A. (2001). Factor analysis with (mixed) observed and latent variables in the exponential family. *Psychometrika*, *66*(4), 515–530.

Widaman, K.F. (1993). Common factor analysis versus principal component analysis: Differential bias in representing model parameters? *Multivariate Behavioral Research*, *28*, 263–311.

Yates, A. (1987). *Multivariate Exploratory Data Analysis: A Perspective on Exploratory Factor Analysis*. State University of New York Press.

Yuan, K. H. & Bentler, P. M. (1998). Normal theory based test statistics in structural equation modelling. *British Journal of Mathematical and Statistical Psychology*, *51*, 289–309.

25 Structural Equation Modeling

Rex B. Kline

Abstract

Structural equation modeling (SEM) is a family of statistical techniques and methods for testing hypotheses about causal effects among observed or proxies for latent variables. There are increasing numbers of SEM studies published in the research literatures of various disciplines, including psychology, education, medicine, management, and ecology, among others. Core types of structural equation models are described, and examples of causal hypotheses that can be tested in SEM are considered. Requirements for reporting the results of SEM analyses and common pitfalls to avoid are reviewed. Finally, an example of evaluating model fit is presented along with computer syntax so that readers can reproduce the results.

Keywords: Structural Equation Modeling; Covariance Structure Analysis; Covariance-Based SEM; Causal Models

Introduction

Structural equation modeling (SEM) is a family of multivariate statistical techniques for estimating presumed causal relations in either observational or experimental studies. The terms path analysis, confirmatory factor analysis, and latent growth curve modeling, among others, all refer to particular types of SEM analyses. It combines aspects of (1) factor analysis, which estimates latent variables (theoretical concepts), given data from their indicators, or observed (manifest) variables (see Chapter 24 in this volume); and (2) regression analysis, that analyzes multiple explanatory variables of the same response (outcome) variable while controlling for intercorrelations among all variables. Equations for multiple variables are simultaneously analyzed such that an outcome variable in one equation can be specified as causal variable in a different equation. Variables can be either observed or latent, and the distinction between observed and latent variables can take account of measurement error. It is also possible to analyze means in SEM, including the comparison of means from independent samples or dependent samples (e.g., repeated measures) on observed variables or proxies for latent variables (Bagozzi & Yi, 2012).

The SEM family is also flexible in that it accommodates analyses that are more exploratory. It can be used in *model generation*, where an initial model is found to be inconsistent with the data and is subsequently modified over a series of follow-up analyses. A second context is testing *alternative models*, where two or more a priori models, comprised of the same variables but specified based on different theories, are all fitted to the same data. A third context is *strictly confirmatory* – a single model that is either retained or rejected based on its correspondence with the data with no further analysis (Jöreskog, 1993). In any context, the goals are to (1) understand patterns of covariances among a set of measured variables, and (2) explain as much of their variance as possible with a statistical model that makes theoretical sense, is parsimonious, and has acceptably close correspondence to the data.

The combination of features just described is relatively unique and probably explains why SEM is being applied in rapidly increasing numbers of studies in disciplines that include psychology, education, medicine, ecology, environmental sciences, commerce, marketing, international business, management, operations research, tourism, and sustainable manufacturing, among others (e.g., Teo, 2010; Thelwall & Wilson, 2016). Many graduate programs now offer courses in SEM, and there are numerous summer schools and seminars on SEM for established researchers around the world. The growing availability of computer tools, some free of charge, has also made SEM more accessible to applied researchers. With no exaggeration, it can be said that SEM is becoming an essential set of statistical techniques.

There are also downsides to the increasing use and popularity of SEM – students or established researchers may be pressured by supervisors or reviewers to use SEM as a cutting-edge method when a simpler statistical technique would do. A related concern is that SEM could be used with little understanding, especially if an emphasis on ease of use of statistical computer tools gives beginners the false impression that SEM is easy (Steiger, 2001). Another problem in many articles is that so much attention is paid to technical aspects of applying SEM that the theoretical sense and meaning of the hypotheses behind the model are neglected. The risk is that, as Tarka (2018, p. 342) put it, "the use of SEM . . . is full of overuse, incorrect interpretation and overinterpretation" – due to apparent unawareness of its potential limitations.

There is also evidence for widespread deficient reporting of results from SEM analyses (e.g., Fan et al., 2016; Shah & Goldstein, 2006). For example, some colleagues and I reviewed a total 144 SEM studies published in 12 top organizational and management journals from 2011 to 2016 (Zhang et al., 2021). Each article was evaluated against criteria that included the clarity of the rationale for using SEM versus alternative techniques, whether hypotheses were tested in a clear and specific order, and whether statistical results were described in sufficient detail. Many shortcomings were apparent: explicit justification for using SEM instead of alternative methods was given in about 40% of the studies; data screening or distributional assumptions were explicitly described in about 20%; and complete details about model fit to the data were reported in about 20% of reviewed studies. That is, the reader of the typical study was not provided with enough information to understand whether the findings actually had any meaningful interpretation. As a reviewer of SEM manuscripts for about 30 different journals, I see these reporting problems all the time.

SEM Families

The term “SEM” does not refer to a single set of statistical techniques or methods. Instead, the three distinct bodies of work listed next, and described afterward, make up what we now refer to as SEM:

1. *Covariance-based SEM* (CB-SEM), also called *covariate structure analysis* or *covariance structure modeling*, is the form of SEM most familiar in psychology, sociology, education, and related disciplines.
2. *Variance-based SEM* (VB-SEM), also known as *composite SEM* or *partial least squares path modeling* (PLS-PM), among other names, is more familiar in research areas on management, organization, and marketing.
3. The *structural causal model* (SCM) originated in Pearl’s (2009) work on Bayesian networks in the 1980–1990s and since extended to the more general problem of causal modeling. The SCM is probably best known in epidemiology, health sciences, and computer science, but that is changing.

All three statistical families just listed owe their origins to pioneering work by the geneticist Sewall Wright (1920), who developed the technique of *path analysis* for testing hypotheses about causal effects among a set of variables. Diagrams for Wright’s models included both observed and latent variables, and they are remarkably similar to modern path diagrams (e.g., Wright, 1920, p. 328). Wright’s work on estimating causal effects using regression methods was introduced to the social and behavioral sciences in the 1960 and 1970s (Tarka, 2018), and the very first computer program for SEM available on mainframe computers, LISREL III (Jöreskog & Sörbom, 1976), combined regression and factor analysis methods, both of which analyze covariances (hence the term “CB-SEM”). All modern SEM computer tools – including the most recent version of LISREL itself (Jöreskog & Sörbom, 2021) – share LISREL III as a forerunner.

The VB-SEM family dates to an approach called “soft modeling,” developed in the 1970s and 1980s by Wold (1982), which estimates latent variables as weighted linear combinations of observed variables, or composites (also called components). Statistical methods for composites are generally less demanding than those that approximate latent variables as common factors (i.e., CB-SEM). For example, composite-based methods may require smaller samples compared with CB-SEM techniques. Analyses of composite models are also less prone to technical problems, such as the failure of iterative estimation to reach a stable solution. Analyses in VB-SEM maximize prediction (i.e., R^2) of outcome variables. In contrast, CB-SEM methods aim to maximize the overall correspondence between model and data; this may not necessarily maximize prediction for individual outcomes. If maximizing prediction is a primary goal, then VB-SEM may be preferred over CB-SEM (Rigdon, 2012).

Pearl’s (2009) SCM corresponds to *non-parametric SEM,* where causal hypotheses are represented in a *directed acyclic graph* (DAG). A DAG is non-parametric because it assumes no particular operationalization for any theoretical variable or any specific functional form of statistical association between variables (e.g., linear versus curvilinear relations for continuous variables). Instead, there are methods in the SCM to analyze a DAG to determine whether it is possible to estimate a target causal effect through the inclusion of other variables as covariates or as instrumental variables, among other possibilities. Thus, a DAG is not a static entity in the SCM; instead, it can be analyzed *with no data whatsoever*, and insights from analyzing the graph can be invaluable in dealing with potential confounding (see Williams et al., 2018 for examples in pediatrics).

It is beyond the scope of this chapter to cover all three SEM families. Instead, just CB-SEM is considered from this point, for two reasons: (1) I would wager that more researchers are familiar with CB-SEM than with VB-SEM or the SCM (Thelwall & Wilson, 2016), and (2) knowing the concepts of CB-SEM provides a strong basis for learning about the other two approaches (Astrachan et al., 2014). This is because knowing something about CB-SEM means that the researcher must (1) understand how concepts are defined, operationalized, and expressed in scores from imperfect observed measures; and (2) comprehend basic principles of regression analysis, factor analysis, and the correct interpretation of standard errors and statistical significance (Kline, 2023; Kühnel, 2001). Ideally, researchers who work with colleagues in other disciplines should know about all three families, but CB-SEM is a good place to start. In the rest of the chapter, the term “SEM” refers to CB-SEM.

SEM Steps and Reporting Standards

Described next are the six basic steps in SEM. They are actually iterative because problems at a particular step may require a return to an earlier step. The context of model generation is assumed.

Step 1: Specification

Specification is the representation of the researcher’s hypotheses as a series of equations or as a model diagram (or both). This involves defining the observed and latent variables and their presumed relations. Outcome (dependent) variables in SEM are referred to as *endogenous variables*. Every endogenous variable has at least one presumed cause among other variables in the model, and error terms that represent unexplained variation are typically associated with each endogenous variable. Depending on the hypothesis, an endogenous variable could be specified as a cause of a different endogenous variable. Endogenous variables, as just described, are *intervening variables* – they are specified as affected by causally prior variables, and in turn they affect other variables further “downstream” in a causal pathway. In contrast, exogenous variables are strictly causal – whatever causes them is not represented in the model.

Whether a variable is endogenous or exogenous is determined solely by the theory being tested. This means that the model is specified *before* the data are collected, and the whole model represents the total set of hypotheses to be evaluated in the analysis. *Specification is the most important step*. This is true because results from the analysis assume that the model is correct. Because the initial model is not always retained, I suggest that, before data collection, researchers make a list of possible modifications that would be *justified according to theory*; that is, prioritize the hypotheses, represent just the very most important ones in the model, and leave the rest for a "backup list." Preregistration of the analysis plan would make strong a statement that changes to the initial model were not made after the examining the data (Nosek et al., 2018).

Step 2: Identification

Identification concerns the issue of whether each model parameter can be expressed as a unique function of the variances, covariances, or means in a *hypothetical* data matrix. It is basically a mathematical proof expressed in *symbolic* form that there is potentially a unique estimator for each effect in the model. *Thus, identification has nothing to do with real data (numbers) or with sample size*. Instead, it is an inherent property of the model, given all the equations that define it. A model that is not identified remains so regardless of both the data matrix and the sample size ($N = 100$, 1,000, etc.) and, thus, must be respecified. An intuitive example follows.

Consider these formulas:

$$\begin{aligned} a + b &= 6 \\ 3a + 3b &= 18 \end{aligned} \tag{25.1}$$

Equation 25.1 is not identified because there is no unique solution (for a, b) that satisfies both formulas. Instead, there are *infinite* solutions, such as (4, 2), (5, 1), and so on. This happens because the second formula in Equation 25.1 is linearly dependent on the first formula – an inherent characteristic. Now consider the formulas listed next where the second is not linearly dependent on the first:

$$\begin{aligned} a + b &= 6 \\ 2a + b &= 10 \end{aligned} \tag{25.2}$$

Equation 25.2 has a single solution – it is (4, 2) – so the whole expression is identified. Structural equation models are typically more complex than Equations 25.1 and 25.2 (i.e., it is often impractical to inspect individual parameters especially in large models). Instead, there are graphical methods and identification heuristics that can determine whether some, but not all, models are identified (Kenny & Milan, 2012). There are also computer tools that analyze diagrams of path models for identification in the SCM approach to SEM (Textor et al., 2020).

Step 3: Measure Selection and Data Collection

Measure selection and data collection are essential activities in most empirical studies. See Kline (2023, ch. 4) and Lang and Little (2018) for how to select measures and deal with potential data-related problems in SEM (e.g., missing data, univariate or multivariate outliers, and extreme collinearity).

Although there have been efforts to make the application of SEM in smaller samples more feasible (Deng et al., 2018), the reality is that SEM is a large-sample technique. Unfortunately, there is no simple answer to the question of how large a sample is needed. This is because sample size requirements vary with model size or type, estimation method, distributional assumptions, and level of measurement for outcome variables, among other considerations. For example, estimation methods in SEM with no distributional assumptions generally need larger samples than methods that assume normal distributions. Larger models with more variables and effects may require larger samples than smaller, simpler models. There is also evidence that sample sizes in many, if not most, published SEM studies are too small in terms of both precision and statistical power (Wolf et al., 2013). As a rule of thumb, $N = 200$ or so might be a reasonable minimum sample size for smaller, more basic models (Barrett, 2007), but 200 is not a magic number. The requirement for large samples complicates replication in SEM, especially when studying rare populations, such as patients with a low-base-rate illness. In this case, it could be challenging to collect a sufficiently large sample for a single analysis, much less twice the number of cases for an additional cross-validation sample, where a model is analyzed in the original sample, and then these analyses are replicated in the second sample of equal size.

Step 4: Analysis

Analysis is carried out using an SEM computer program to fit the model to the data. A few things take place at this step. First, (a) evaluate model fit to determine how well the model fits the data. Often, the initial model does not adequately explain the data; if so, skip the rest of this step and go to step 5. Otherwise, next (b) interpret the parameter estimates, and (c) consider *equivalent models* that fit the data *exactly* as well as the researcher's model but feature contradictory hypotheses about causation among the same variables (Henley et al., 2006). An example is presented later, but the failure to acknowledge equivalent models is a widespread problem in SEM studies that is also a form of confirmation bias.

Step 5: Respecification

In this step, the initial model is altered and fitted to the same data, *but any respecified model must be theoretically justified* (i.e., consult the backup list mentioned earlier). If there is no such justification, it may be better – and more honest, too – to retain no model (Hayduk, 2014), especially compared with making changes solely to improve the fit of the model in a particular sample. The problem is that post doc, data-driven respecification can lead to a model that does not replicate because it capitalizes so strongly on sample-specific variation.

Step 6: Reporting

This is the written summary of the results. If a model is retained, describe both global fit and local fit. *Global fit* concerns the overall or average match between the model and the data matrix. Just as averages do not indicate variability, models in SEM with apparently satisfactory global fit can have problematic *local fit*; this is measured by residuals calculated for every pair of measured variables (see Tomarken & Waller, 2003 for examples). Residuals in SEM concern differences between observed (i.e., in the data) versus predicted (i.e., from the model) covariances or correlations, and as absolute residuals increase in size, local fit becomes worse. The analogy in regression is the difference between R^2 – overall predictive power (global fit) – and regression residuals – differences between observed and predicted scores. Aberrant patterns of regression residuals indicate a problem in the analysis even if the value of R^2 is reasonably high. Just as reports about regression results with no mention of the residuals are incomplete, so too are reports in SEM in which only global model fit is described. For an example of full reporting on residuals in SEM, see Sauvé et al. (2019, Appendix A).

Reporting about both global and local model fit is part of journal article reporting standards for SEM studies by the American Psychological Association (Appelbaum et al., 2018) and this is based on earlier standards for SEM by Hoyle and Isherwood (2013) for the journal *Archives of Scientific Psychology*. Reporting standards also call on researchers to

(a) outline how the sample size was determined, such as through power analysis
(b) give a full account of model specification, including the rationale for hypotheses about the directionality of causal effects (i.e., X causes Y and not the reverse), all in the context of relevant theory
(c) explain the bases for respecification of an initial model and whether respecifications were a priori or post hoc
(d) interpret statistical results according to evidence-based criteria
(e) justify the preference for any retained model over equivalent models that explain the data just as well
(f) report the unstandardized solution with standard errors and the standardized solution
(g) report sufficient summary statistics to allow secondary analysis or make the raw data file available.

SEM Computer Programs

In the late 1970s, LISREL was among a small number of computer tools for SEM, but today there are many options for SEM software, both commercial and freely available. Free software packages for SEM include lavaan (Rosseel et al., 2022) and OpenMx (Boker et al., 2022) for the R computing environment. There are also R packages for conducting specialized types of SEM analyses, such as semTools for simulation and power analysis (Jorgensen et al., 2022). Other free options that do not involve R include JASP, an integrated, open-source application with capabilities for traditional (frequentist) and Bayesian analyses (including SEM; JASP Team, 2022), and Ωnyx (pronounced “onyx”), which features a drawing editor where the user specifies the model and controls the analysis by drawing the model on the computer screen (von Oertzen et al., 2015).

Free-standing commercial products for SEM analyses include Amos, EQS, Mplus, and LISREL (respectively, Arbuckle, 2021; Bentler & Wu, 2020; Müthen & Müthen, 1998–2017; Jöreskog & Sörbom, 2021). Some widely used software for general statistical analyses have procedures, functions, or commands for SEM. Examples include the sem command in Stata (StataCorp, 1985–2021) and the CALIS procedure in SAS/STAT (SAS Institute, 2021). Some universities and research centers have site licenses for commercial SEM software that allow free use by researchers and students, but individual licenses can be relatively expensive. Commercial products have the advantage of complete manuals with many analysis examples or data sets, if cost is no problem; otherwise, free SEM software (e.g., lavaan) is nearly as capable as commercial products.

Core Types of Models

Described next are three core types of models in SEM with examples of each from actual studies. All example models are identified. Presented in Figure 25.1 is the *manifest-variable (classical) path model* analyzed by Yamaga et al. (2013). Such models feature *single-indicator measurement*, where each construct is measured by a single observed variable. For all examples,

(1) observed variables are represented with squares or rectangles
(2) latent variables are depicted with circles or ovals
(3) lines with single arrowheads point from presumed causes to endogenous variables
(4) presumed covariances between measured variables are represented as curved lines with arrowheads at each end.

It is also common in model diagrams to represent error terms for endogenous variables, but there is no standard symbolism for doing so. Perhaps the most basic symbol is a line with a single arrowhead oriented at a 45-degree angle that points to each outcome (e.g., ↙; see Figure 25.1), but McDonald and Ho (2002) describe additional ways to graphically represent error terms in diagrams of structural equation models.

In a sample of 166 edentulous (toothless) dental patients who presented themselves for complete denture therapy, Yamaga et al. (2013) measured the integrity of mandibular ridge form (lower jaw bone formation), retention and stability of mandibular complete denture, jaw relation (whether the cusps of opposing teeth on the lower and upper [maxillary] jaws correctly interlock), perceived chewing ability (mastication), satisfaction with exiting complete dentures, and extent of oral health problems. Their path model in Figure 25.1 represents the hypotheses that

(1) ridge form, retention, and stability all co-vary and also directly affect jaw relation
(2) jaw relation, in turn, is a direct cause of both mastication and denture satisfaction
(3) mastication is also caused by stability, and satisfaction is also affected by both ridge form and mastication
(4) oral health problems are directly affected by both mastication and satisfaction.

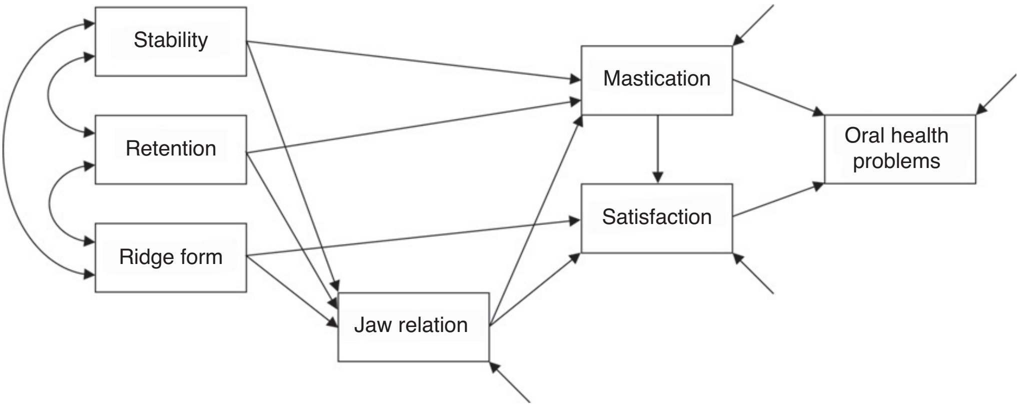


Figure 25.1 *Example of a manifest-variable path model analyzed by Yamaga et al. (2013).*

The variables jaw relation, mastication, and denture satisfaction in Figure 25.1 are specified as intervening variables that "absorb" effects from prior causal variables and "transmit" those effects to subsequent outcomes. For example, the indirect pathway

stability → jaw relation → mastication

represents the hypothesis that stability of mandibular complete denture affects jaw relation, which, in turn, impacts mastication. Indirect effects are part of the concept of mediation, *but the two are not synonymous*. This is because *mediation* is the strong causal hypothesis that one variable (stability) causes *changes* in another variable (jaw relation), which leads to *changes* in an outcome (mastication; Little, 2013). The emphasis on "changes," in the definition just stated, highlights the requirement for *time precedence* – measurement of presumed causes before their outcomes. With no time precedence, it is difficult to interpret estimates for indirect effects as evidence for mediation (Pek & Hoyle, 2016). Yamaga et al.'s (2013) design was cross-sectional – all variables were measured at the same occasion (see Chapter 13 in this volume) – so the term "mediation" does not automatically apply to any of the indirect causal pathways in Figure 25.1.

Especially in cross-sectional designs, which have no inherent support for causal inference, *directionalities of causal effects in SEM are assumed, not tested*. This is because there is little, if anything, from analysis that could either disconfirm or verify hypotheses about causal priority. For example, outcomes of significance testing for the *path coefficient* of $X \rightarrow Y$, a presumed direct effect of X on Y, could fail to be significant in a small sample due to insufficient power. The phenomenon of equivalent models with the opposite specification – $Y \rightarrow X$ – which fit the data just as well as the original model, discounts the possibility that significant path coefficients prove causation. This is why it is critical to provide clear and reasoned justifications for directionality specifications, especially in cross-sectional designs. Thus, SEM is not a technique for causal discovery. This means that, if given a true model, SEM could be applied to estimate the magnitudes of causal effects represented in the model. However, this is not how SEM is typically used; instead, a causal model is *hypothesized*, and the model is fitted to sample data *assuming* that all its specifications are correct.

Other assumptions of the path model in Figure 25.1 are briefly summarized next: (1) Score reliabilities on the exogenous variables (stability, retention, and ridge form) are perfect – $r_{XX} = 1.0$. Exogenous variables in path models, as in the figure, have no error terms, so there is no "room" for measurement error in these variables. This requirement does not apply to endogenous variables (e.g., jaw relation) that have error terms that absorb measurement error. (2) There are no unmeasured common causes, or confounders, for any pair of variables in the model. This assumption is required because the omission of confounders can seriously bias values of coefficients in both regression analysis and SEM (Cohen et al., 2003). (3) The error terms in Figure 25.1 are independent; that implies all unmeasured causes of endogenous variables are all pairwise uncorrelated and also with all three exogenous variables. Altogether, these assumptions are very demanding. Results from analysis of the model in Figure 25.1 are described in the last section of this chapter.

Figure 25.2 shows a *confirmatory factor analysis* (CFA) model analyzed by Filippetti and Krumm (2020), who administered five performance tasks, hypothesized to reflect two dimensions of cognitive flexibility, to 112 children aged 8–12 years. These domains included reactive flexibility – the capability to modify behavior – and spontaneous flexibility– the ability to generate novel responses. The two language-based fluency tasks in the figure involve asking examinees to say as many words as possible for two categories (e.g., animals; semantic fluency) or starting with a specific letter (e.g., S; phonetic fluency) for 60 seconds. The pattern fluency task measures the ability to produce unique geometric designs within a time limit. The three tasks just described are specified as indicators of spontaneous flexibility; this is represented in Figure 25.2 with the symbol for a latent variable – an oval (circles can also designate latent variables in model diagrams). Indicators in CFA models have error terms that capture random measurement error in the observed variables. Thus, it is *not* assumed in CFA that the scores are perfectly precise.

The remaining two observed variables in Figure 25.2 are specified as indicators of reactive flexibility. These tasks include a computerized card-sorting task, where examinees are asked to match geometric patterns. Because they are told only whether their responses are correct or incorrect, examinees must infer the matching rules. The trail-making task requires examinees to draw lines in an alternating series of numbers and letters in sequential order (e.g., 1, A, 2, B, and so on). Numerals (e.g., 1) that appear in the figure next to certain direct effects (one per factor) are *scaling constants* that specify metrics for the factors. For example, the specification in the figure

$$\text{reactive flexibility} \rightarrow \text{card sorting} = 1$$

assigns a scale to the reactive flexibility factor. That scale corresponds to variation in the card-sorting task that is explained by the factor it is presumed to measure. It is

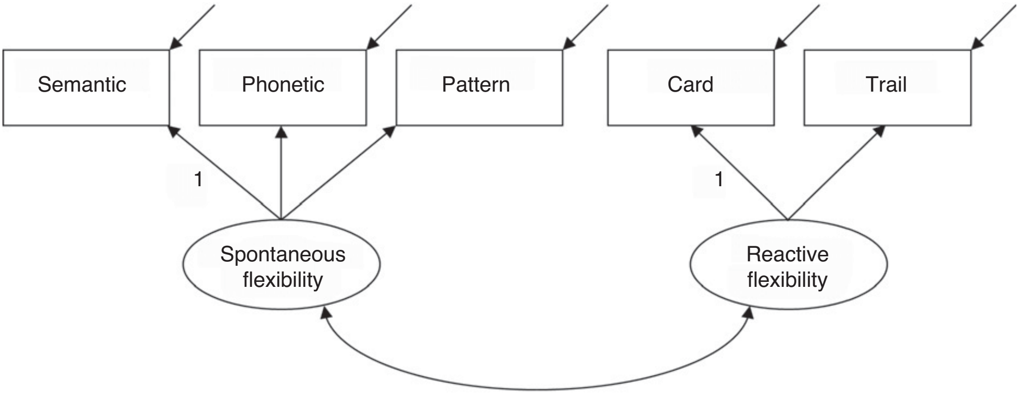


Figure 25.2 *Example of a CFA model analyzed by Filippetti and Krumm (2020).*

usually arbitrary which direct effect is so specified, but latent variables must be scaled before the computer can derive statistical estimates about them (see Brown, 2015, for discussion of other options to scale factors).

The symbol for a covariance that connects the spontaneous flexibility and reactive flexibility factors in Figure 25.2 instructs the computer to estimate their covariance (unstandardized) or correlation (standardized), given the model and data. It is often reasonable to assume that hypothetical constructs are related, such as cognitive ability factors (e.g., verbal, visual–spatial, memory), and covariances between all pairs of factors are routinely estimated in CFA. If two factors are believed to be independent, their covariance can be specified as zero to test this hypothesis (see Brown, 2015 for examples). In addition to the two-factor, five-indicator structure represented in Figure 25.2, the model also assumes that (1) omitted causes of the indicators are unrelated to the factors, and (2) omitted causes for each indicators have no overlap with those for all other indicators.

It is important, in any method of factor analysis, to avoid the *naming fallacy*; just because a factor is named does not mean that the corresponding hypothetical construct is understood or even correctly labeled. For instance, the label "reactive flexibility" in Figure 25.2 does not preclude other interpretations of what the card-sorting and trail-making tasks measure (e.g., abstract reasoning or visual analysis). Factor labels are conveniences that are more "reader friendly" than abstract symbols, but they are not substitutes for critical thinking (Kline, 2023). Another potential error is *reification* – the false belief that a factor *must* correspond to something in the real world. Factors are statistical abstractions from observed measures, and whether such abstractions describe any tangible entity, dimension, or process is an open question (Rigdon, 2012).

Figure 25.3 shows a *structural regression (SR) model*, also called a *latent-variable path model* or *full-LISREL model* because LISREL was one of the first computer programs to analyze such models. The SR model in the figure was analyzed by Recio et al. (2013), who administered measures of executive function – cognitive processes needed for monitoring and control of behavior (e.g., attentional focus) – and

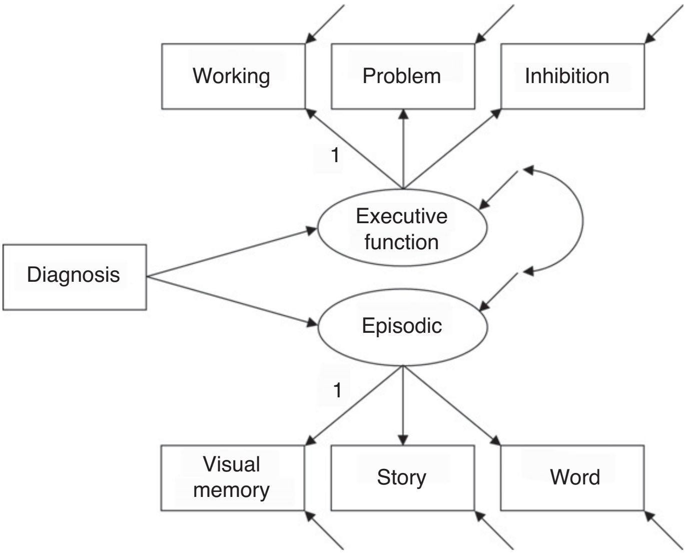


Figure 25.3 *Example of a SR model analyzed by Recio et al. (2013).*

measures of episodic memory – recall of visual or auditory stimuli – within samples of patients with Parkinson's disease and neurologically healthy adults matched for age and level of education. It is worth noting that the group sizes were, I believe, too small – a total of 23 patients and 18 control cases – for precise estimation of the model in Figure 25.3. A more reasonable group size would be $n \geq 100$, but even that number may be inadequate for sufficient statistical power.

The measurement part of the model corresponds to the two factors, each with three indicators: working memory, problem solving, and inhibition for the executive function factor, and tests of visual, story, and word recall for the episodic memory factor. Both factors just mentioned are specified as outcomes of the dichotomous variable of diagnosis, which specifies membership in either the Parkinson's disease group or the control group. Because the factors are endogenous in Figure 25.3, they each have error terms that represent variation not explained by diagnosis. In contrast, factors in CFA models are exogenous and do not have error terms (cf. Figure 25.2).

The curved line with arrowheads at each end in Figure 25.3 represents an *error covariance* in the unstandardized solution or an *error correlation* in the standardized solution. This specification instructs the computer to estimate the association between the executive function and episodic memory factors *after controlling for diagnosis*. Here it makes sense that the two cognitive factors would be related above and beyond the distinction between Parkinson's disease and control cases. Other valid reasons to specify correlated error terms in SEM include autocorrelation among variables in longitudinal designs, common response sets (systematic difference in how participants respond to questions regardless of item content), and shared stimuli over tasks (Westfall et al., 2012). Each error correlation added to a model makes it more complex and generally improves fit. A concern is that error correlations are added mainly to enhance fit without substantive reasons. As with any other model specification, the inclusion of correlated errors requires justification.

Example SEM Analysis and Reporting Recommendations

In their original analysis of the path model in Figure 25.1, Yamaga et al. (2013, p. 14) reported sufficient summary statistics for their raw data – correlations and standard deviations – to allow other researchers to reproduce their results in a secondary analysis (with slight rounding errors). Doing so is a best practice both in SEM and other types of quantitative studies (Appelbaum et al., 2018). This is because, even with no access to the raw data, other researchers can independently verify the original analyses or test hypotheses not considered by the authors of the original work. There are some types of SEM analyses that require raw data files. Examples include the analysis of continuous variables with methods that adjust for severely non-normal distributions or the analysis of ordinal data (see Kline, 2023, for more information), but summary statistics are all that's needed in this example. Listed in the appendix at the end of this chapter is syntax for lavaan that fits the model in Figure 25.1 to summary statistics reported by Yamaga et al. (2013) for $N = 166$ cases. This syntax can be executed in R after installing the lavaan package – install.packages("lavaan", dependencies = TRUE). The output file will contain all the results described next. The estimation method is default maximum likelihood and assumes normal distributions.

Listed next is a suggested structure for the results section that is also consistent with reporting standards for SEM (Appelbaum et al., 2018). It is assumed that the theoretical rationale for model specification is outlined earlier in the manuscript:

(1) Explicitly tabulate numbers of observations, free model parameters, and model degrees of freedom.
(2) Report results about both global model fit and local model fit, or the residuals.
(3) Justify the decision to either retain the model as initially specified, reject the model before the analysis enters a respecification phase, or reject the model with no further changes nor analyses. If a respecified model is retained, state the rationale for any modifications to the initial model, including whether respecification was mainly a priori or empirical.
(4) If a model is retained, then (a) report the unstandardized parameter estimates with standard errors and the standardized solution. Also, (b) directly acknowledge the existence of equivalent models, generate at least a few examples, and argue why the retained model is preferable to any equivalent version with exactly the same fit to the data.

Model Degrees of Freedom

The *model degrees of freedom*, df_{M}, is the difference between the number of observations and the number of free model parameters. The number of *observations* for continuous variables, when means are not analyzed (as in this example), equals $v\ (v + 1)/2$, where v is the number of observed variables. For example, $v = 7$ in Figure 25.1, so the number of observations is 7(8)/2, or 28; this equals the number of elements in the covariance matrix generated by the descriptive statistics in Yamaga et al. (2013, p. 14) in lower diagonal form, where redundant values above the diagonal are eliminated (see the appendix).

A *free parameter* is estimated by the computer with the sample data. Free parameters when means are not analyzed include (1) variances and covariances of exogenous variables, (2) direct effects on endogenous variables from other variables in the model (but not error terms), and (3) the variance of each error term (Kline, 2023). In Figure 25.1, there are three exogenous variables with a covariance between each pair, so the total number of variances and covariances here is 3 + 3 = 6. There are four endogenous variables in the figure with a total of 11 direct effects on them from other variables. Each endogenous variable has an error term; a total of four error variances must be estimated by the computer. Thus, the total number of free parameters is

$$6 + 11 + 4 = 21 \text{ so } df_{\mathrm{M}} = 28 - 21 = 7.$$

Models with no degrees of freedom ($df_{\mathrm{M}} = 0$) will perfectly fit the data. This is because such models are as complex, in terms of free parameters versus observations, as the data they are supposed to explain. Models where $df_{\mathrm{M}} = 0$ test no particular hypothesis and, thus, are rarely of interest. Positive degrees of freedom ($df_{\mathrm{M}} > 0$) allow for the *possibility* of discrepancies between model and data – imperfect fit. A key question in the analysis for models with $df_{\mathrm{M}} > 0$ is whether expected differences between model and data are so great that the model should be rejected. Thus, $df_{\mathrm{M}} > 0$ is an effective requirement in SEM, and there is a preference for models with greater degrees of freedom or models that are more parsimonious based on df_{M} (Raykov & Marcoulides, 2006). Models with negative degrees of freedom ($df_{\mathrm{M}} < 0$) are not identified and must be respecified so that $df_{\mathrm{M}} \geq 0$ before they can be analyzed.

Global Fit

There are two kinds of global fit statistics in SEM: model test statistics (i.e., significance tests) and approximate fit indexes; these are not significance tests. The most widely reported test statistic is the *model chi-square* with its degrees of freedom, df_{M}. The statistic is designated here as chi$_{\mathrm{M}}$. It tests the null hypothesis that the researcher's model perfectly fits the *population* data matrix. The value of chi$_{\mathrm{M}}$ equals the product of sample size (N) and the degree of difference between the sample data matrix and associations for the same variables predicted by the researcher's model. If chi$_{\mathrm{M}} = 0$, the model perfectly fits the sample data matrix. As model–data discrepancies increase, the value of chi$_{\mathrm{M}}$ increases, too.

If $\text{chi}_{\text{M}} > 0$ and its p-value is less than α, the criterion level of statistical significance, then (1) the null hypothesis of perfect fit is rejected, and (2) the model *fails* the chi-square test. Suppose that $\text{chi}_{\text{M}} = 12.50$ for a model where $df_{\text{M}} = 5$. The p-value for this result is 0.029. If $\alpha = 0.05$, the model fails the chi-square test because $p < \alpha$. This means that the difference between the data matrix and the predicted matrix is significant at the 0.05 level. *Passing* the model chi-square test happens whenever $p \geq \alpha$, such as chi_{M} (5) = 10.50, $p = 0.062$ (when testing at the 0.05 level). *But passing the chi-square test does not automatically mean that the model also has satisfactory local fit*. It can and does happen, especially in samples that are not large and where the power of the chi-square to detect appreciable model–data discrepancies is low, that passing models have poor local fit; that is, the residuals are problematic. *Such models should not be retained even though they passed the chi-square test*. Likewise, it can happen, in very large samples, that a model fails the chi-square test, but the residuals indicate trivial discrepancies in local fit. In this case, the researcher might reasonably argue to retain the model, given satisfactory residuals. In fact, some researchers used to divide chi_{M} by df_{M} to reduce its sensitivity to sample size, but (1) df_{M} has nothing to do with sample size, and (2) there are never any specific values of $\text{chi}_{\text{M}}/df_{\text{M}}$ that indicate "good" fit (e.g., < 3.0, 5.0, or some other value). Therefore, I do not recommend it.

A widespread but poor practice in published SEM studies occurs when (1) the model fails the chi-square test but (2) the researcher automatically dismisses this result because "the model chi-square is affected by sample size" or some such rationale that is actually false; N affects chi_{M} *only when the model is wrong* (Hayduk, 2014). Failing the chi-square test should be interpreted as indicating covariance evidence against the model, and that failure should be thoroughly diagnosed (i.e., inspect the residuals). Passing the chi-square test should also be followed by careful inspection of the residuals *because the details of fit are in the residuals*.

Approximate fit indexes are continuous measures of model–data discrepancy. Some approximate fit indexes are scaled like chi_{M} – a value of zero is the best result concerning model fit. Others have more-or-less standardized metrics from 0 to 1.0, where 1.0 is the best result. Dozens of approximate fit indexes have been described in the literature, and output for some SEM computer tools includes values for rather lengthy lists of such indexes. A problem with all approximate fit indexes is that there is little correspondence between their numerical values and types or seriousness of specification error (Hayduk, 2014). The same thing is true about chi_{M}, its p-values, and the residuals. One reason is equivalent models that have identical values of all global fit statistics *and* residuals even though they represent contradictory sets of hypotheses.

An issue with approximate fit indexes is overreliance on now-discredited fixed thresholds that supposedly indicate whether model–data correspondence is "good." An example of a "golden rule" for the hypothetical "ABC" global fit statistic is, "if ABC > 0.95, then model fit is good." Such thresholds date from computer simulation studies in the 1980s and 1990s about a very narrow range of models, but subsequent results indicated that these fixed thresholds do not always apply to other kinds of models or data (Barrett, 2007). For example, fixed thresholds were originally developed for models with continuous endogenous variables, but they are not accurate for models with ordinal endogenous variables (Xia & Yang, 2019). There is no problem with reporting values of approximate fit indexes, *but there are no magic cutting points that somehow differentiate between models with "good" versus "poor" fit*, especially if the researcher does not also look to the residuals for more detailed information about model fit.

Listed next is what I believe is a minimal set of approximate fit indexes that should be reported in most analyses (Kline, 2023, ch. 10). Mulaik (2009) describes additional indexes for special contexts, but I think many reviewers of submissions to journals would expect to see the minimal set:

(1) The *Steiger–Lind root mean square error of approximation* (RMSEA) and its 90% confidence interval (CI); in contrast to chi$_{\mathrm{M}}$, which measures departure from perfect fit, the RMSEA measures departure from approximate fit in a correlation metric that also controls for *N* and df_{M}. *Approximate fit* means that chi$_{\mathrm{M}}$ does not exceed its expected value, df_{M}, over random samples when the model is true in the population. The best result is RMSEA = 0.

(2) The *Bentler comparative fit index* (CFI) compares the relative departures from approximate fit of the researcher's model compared with that of a baseline model in a standardized metric in which CFI = 1.0 is the best result.

(3) The *standardized root mean square residual* (SRMR) approximately measures the average absolute discrepancy between sample correlations and those predicted by the researcher's model for every pair of measured variables. The best result is SRMR = 0.

Values of global fit statistics computed in lavaan for the example analysis are:

$$\begin{aligned}&\mathrm{chi}_{\mathrm{M}}(7) = 7.320,\ p = 0.396\\ &\mathrm{RMSEA} = 0.017, 90\%\ \mathrm{CI}[0, 0.098]\\ &\mathrm{CFI} = 0.999, \mathrm{SRMR} = 0.042\end{aligned}$$

The model passes the chi-square test at the 0.05 level. However, the sample size is small, and the power of the chi-square test, in this analysis for *N* = 166 estimated in semTools, is only 0.18. This means that, if the model does *not* have perfect fit in the population, there is only a likelihood of 0.18 that this status will be detected in the chi-square test. Although RMSEA = 0.017 is not a terrible result, the upper bound of its 90% CI, or 0.098, is very close to 0.10, or so high that it signals *possible* ill fit at the level of the residuals. The result CFI = 0.999 is not alarming, and it says that the model in Figure 25.1 reduces the relative amount of departure from approximate fit by nearly 100% compared with a null model that assumes the endogenous variables are independent of each other and the exogenous variables. The result for the SRMR says that the average absolute difference between sample correlations and those predicted by the model is about 0.042; this is not a terrible result, but it masks problems at the level of the residuals.

Local Fit

Yamaga et al. (2013) did not describe the residuals in their original analysis, but we consider these results computed in lavaan for the same model and data. Reported in the top part of Table 25.1 are *correlation residuals* – differences between observed and predicted correlations for every pair of observed variables. They are continuous measures of local model–data discrepancies, and their values are relatively unaffected by sample size. Absolute correlation residuals > 0.10 signal a potential problem (Kline, 2023; Tabachnick & Fidell, 2013). It is hard to say exactly how many absolute correlation residuals ≥ 0.10 is too many, but the more there are, the worse the local fit. In Table 25.1, two absolute correlations shown in boldface exceed 0.10. For example, the correlation residual for the variables retention and oral health problems is −0.116. The sample correlation is −0.309 (Yamaga et al., 2013, p. 714), so the model underpredicts their association by −0.116 (i.e., the predicted correlation is −0.193). The path model in Figure 25.1 has no direct effect between these two variables, so perhaps that specification is an error (among other possibilities). The absolute correlation residual for ridge form and oral health problems, −0.114, is also relatively high (see Table 25.1).

The bottom part of Table 25.11 shows the *standardized residuals* – significance tests in the form of normal deviates (z) of the corresponding *covariance (raw, unstandardized) residuals*, or differences between sample and predicted covariances. Because covariances reflect the raw score metrics of both variables, it can be difficult to interpret the meaning of covariance residuals. Standardized residuals are more straightforward in their interpretation: If $z > 1.96$ in absolute value, then the corresponding covariance residual differs significantly from zero at the 0.05 level. In small samples, the power of standardized residuals is probably low. Nevertheless, the standardized residual for the pair retention and oral health problems (−2.214) is significant at the 0.05 level, and the result for the pair ridge form and oral health problems (−1.814) is nearly so (Table 25.1). Overall, there are signs of problematic fit at the level of the residuals, and any enthusiasm about global model fit should be

Table 25.1 *Correlation residuals and standardized residuals for a path model of denture satisfaction and oral health*

Variable	1	2	3	4	5	6	7
			Correlation residuals				
1. Jaw relation	0						
2. Mastication	0	0					
3. Satisfaction	0	0.009	0.007				
4. Oral health	−0.038	−0.004	−0.007	0.004			
5. Stability	0	0	0.051	−0.071	0		
6. Retention	0	0	0.071	**−0.116**	0	0	
7. Ridge form	0	0.080	0.033	**−0.114**	0	0	0
			Standardized residuals				
1. Jaw relation	0						
2. Mastication	0	0					
3. Satisfaction	0	1.195	1.195				
4. Oral health	−0.805	−1.195	−1.195	1.195			
5. Stability	0	0	0.892	−1.314	0		
6. Retention	0	0	1.228	**−2.124**	0	0	
7. Ridge form	0	1.195	1.195	−1.814	0	0	0

tempered here by knowledge of relatively poor explanatory power for certain pairs of variables in the model at the level of the residuals.

Equivalent Models

Next, we consider equivalent models. Figure 25.4(a) shows is the original Yamaga et al. (2013) path model for which chi_{M} (7) = 7.320. The other three models in the figure are equivalent versions generated by the *replacing rules*; they permit the substitution or reversal of certain paths *without affecting model fit* (Williams, 2012). For Figures 25.4 (a)–4(d), chi_{M} (7) = 7.320, and values of all other fit global fit statistics and residuals are exactly equal, but the equivalent models in Figures 25.4(b)–4(d) make opposing causal claims. For example,

(1) the status of the ridge form variable in Figure 25.4(b) is changed from exogenous, or causal in the original model, to endogenous, or an outcome in this equivalent version
(2) the direct causal effect between the ridge form and jaw relation variables is revered in Figure 25.4(c) compared with the original model
(3) the stability variable is specified as endogenous in Figure 25.4(d) – stability was exogenous in the original model
(4) the direct effect between stability and jaw relation is reversed in Figure 25.4(d) compared with Figure 25.4(a).

More equivalent versions of the original path model could be generated, so the variations in Figure 25.4 are not exhaustive. Yamaga et al. (2013) did not address the issue of equivalent models, a common shortcoming in SEM studies. A best practice would be for researchers to acknowledge the existence of at least a few plausible equivalent models and then argue why the original version is preferred. For example, Kale et al. (2000) retained a model of conflict resolution and relational capital, generated an equivalent version with identical fit, and gave arguments for their preferred model over the equivalent version. This level of transparency in SEM is commendable but rare.

Summary

The SEM family of techniques is flexible, used in many different areas, and can test a wide range of hypotheses about observed or latent variables. However, there are downsides to its increasing use in the social and behavioral sciences. This is especially true regarding incomplete reporting of the results, such as neglecting to describe full details about model fit. Respecting formal reporting standards for SEM would help to reduce incomplete reporting. Another common shortcoming is the failure to acknowledge the existence of equivalent models that explain the data just as well as the researcher's model. There are many additional kinds of models and analyses that are possible in SEM, but all require good judgment in their application and open, transparent reporting.

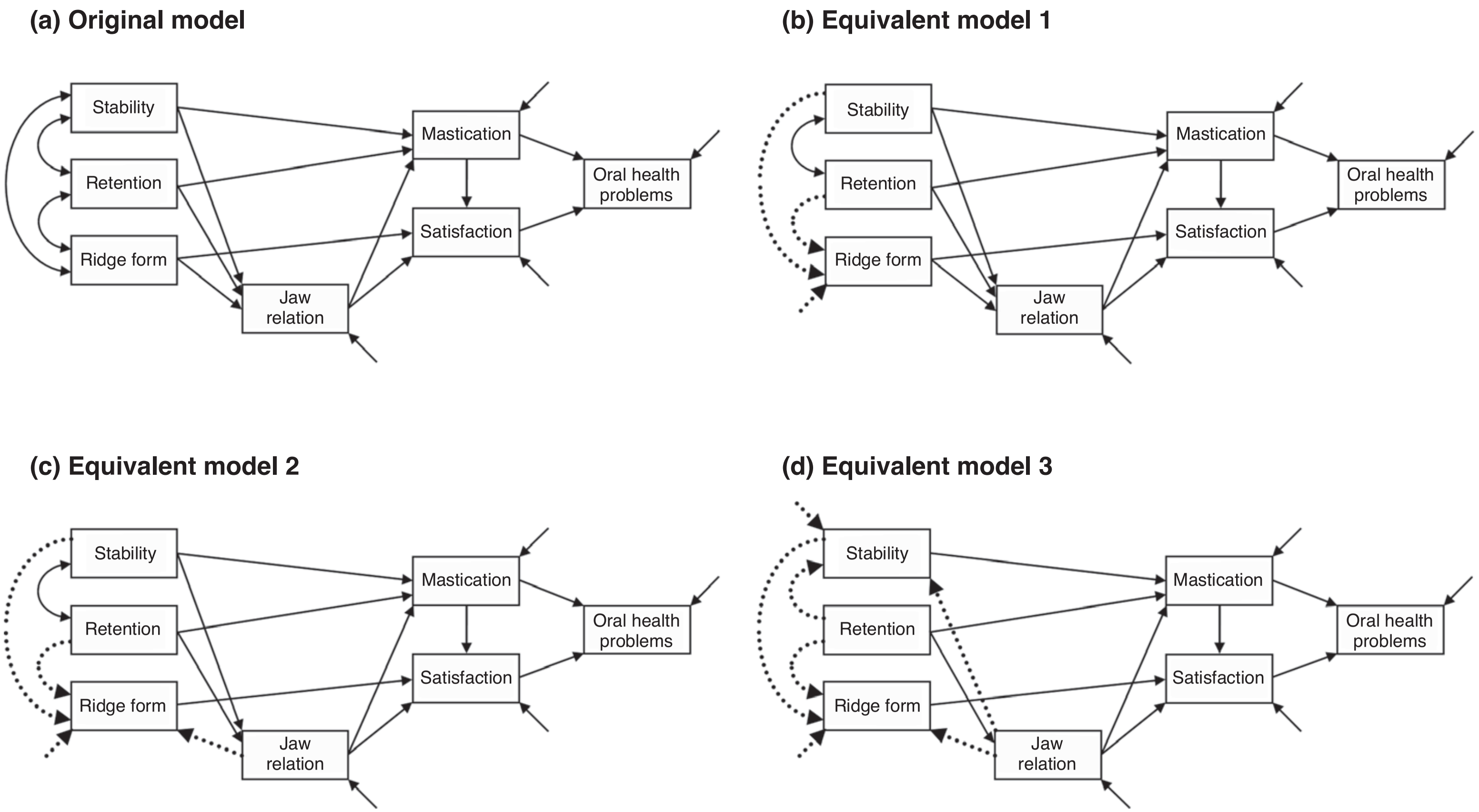


Figure 25.4 *Original Yamaga et al. (2013) path model (a) and equivalent versions (b–d) all with identical fit to the data; dotted lines changed causal status relative to original model.*

Appendix

Syntax in lavaan for specifying and analyzing the example path model

```
# yamaga et al. (2013) path model
date()
options(”width” = 130)
library(lavaan)
library(semTools)
citation(”lavaan”, auto = TRUE)
citation(”semTools”, auto = TRUE)
# input data (covariances)
yamagaLower.cov <-'
  1.2769000
  0.1957612 0.5041000
  0.2490746 0.5287512 1.1449000
  0.3366496 0.2143064 0.3166772 0.9604000
  6.9383130 6.1569780 7.8776610 9.0972420 846.810000
  9.1743570 5.8470630 8.6872230 12.4053300 439.494390 846.810000
-4.4097120–2.9956320–4.7610720 -4.9815360–266.928480 -265.252320 207.360000 '
# add variable names
yamaga.cov <- getCov(yamagaLower.cov, names = c(“ridgeform”,"stability”,
  “retention”,"jawrelation”,"mastication”,“satisfaction”,
  ”oralhealth”))
# display covariances
yamaga.cov
# specify path model
  yamaga.model <-'
  jawrelation ~ stability + retention + ridgeform
  mastication ~ stability + retention + jawrelation
  satisfaction ~ ridgeform + jawrelation + mastication
  oralhealth ~ satisfaction + mastication '
# fit model to data, N = 166
yamaga.lavaan <- sem (yamaga.model, sample.cov = yamaga.cov,
  sample.nobs = 166)
summary(yamaga.lavaan, fit.measures = TRUE, standardized = TRUE,
  rsquare = TRUE)
# predicted covariance matrix
fitted(yamaga.lavaan)
# unstandardized, standardized, and correlation residuals
residuals(yamaga.lavaan, type = “raw”)
residuals(yamaga.lavaan, type = “standardized”)
residuals(yamaga.lavaan, type = "cor.bentler")
# power of the chi-square test
findRMSEApower(0, .05, 7, 166, .05, 1)
```

References

Appelbaum, M., Cooper, H., Kline, R. B., et al. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board Task Force report. *American Psychologist*, *73*(1), 3–25. https://doi.org/10.1037/amp0000191

Arbuckle, J. L. (2021). *IBM SPSS Amos 28 User's Guide*. Amos Development Corporation.

Astrachan, C. B., Patel, V. K., & Wanzenried, G. (2014). A comparative study of CB-SEM and PLS-SEM for theory development in family firm research. *Journal of Family Business Strategy*, *5*(1), 116–128. https://doi.org/10.1016/j.jfbs.2013.12.002

Bagozzi, R. P. & Yi, Y. (2012). Specification, evaluation, and interpretation of structural equation models. *Journal of the Academy of Marketing Science*, *40*(1) 8–34. https://doi.org/10.1007/s11747-011-0278-x

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, *42*(5), 815–824. https://doi.org/10.1016/j.paid.2006.09.018

Bentler, P. M. & Wu, E. J. C. (2020). EQS 6.4 for Windows [Computer software]. Available at: https://mvsoft.com/.

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research*, 2nd ed. Guilford Press.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd ed. Erlbaum.

Deng, L., Yang, M., & Marcoulides, K. M. (2018). Structural equation modeling with many variables: A systematic review of issues and developments. *Frontiers in Psychology*, *9*, Article 580. https://doi.org/10.3389/fpsyg.2018.00580

Fan, Y., Chen, J., Shirkey, G., et al. (2016). Applications of structural equation modeling (SEM) in ecological studies: An updated review. *Ecological Processes*, *5*(1), Article 19. https://doi.org/10.1186/s13717-016-0063-3

Filippetti, V. A. & Krumm, G. (2020). A hierarchical model of cognitive flexibility in children: Extending the relationship between flexibility, creativity and academic achievement. *Child Neuropsychology*, *26*(6), 770–800. https://doi.org/10.1080/09297049.2019.1711034

Hayduk, L. A. (2014). Shame for disrespecting evidence: The personal consequences of insufficient respect for structural equation model testing. *Medical Research Methodology*, *14*(1), Article 124. https://doi.org/10.1186/1471-2288-14-124

Henley, A. B., Shook, C. L., & Peterson, M. (2006). The presence of equivalent models in strategic management research using structural equation modeling: Assessing and addressing the problem. *Organizational Research Methods*, *9*(4), 516–535. https://doi.org/10.1177/1094428106290195

Hoyle, R. H. & Isherwood, J. C. (2013). Reporting results from structural equation modeling analyses in *Archives of Scientific Psychology*. *Archives of Scientific Psychology*, 1, 14–22. https://doi.org/10.1037/arc0000004

JASP Team (2022). JASP (Version 0.16.1) [Computer software]. Available at: https://jasp-stats.org/

Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & J. S. Lang (eds.), *Testing Structural Equation Models* (pp. 294–316). SAGE Publications.

Jöreskog, K. G. & Sörbom, D. (1976). *LISREL III: Estimation of Linear Structural Equation Systems by Maximum Likelihood Methods*. National Educational Resources.

Jöreskog, K. G. & Sörbom, D. (2021). LISREL 11 for Windows [Computer software]. Available at: https://ssicentral.com/.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2022). semTools: Useful tools for structural equation modeling (R package 0.5-6). Available at: https://CRAN.R-project.org/package=semTools.

Kale, P., Singh, H., & Perlmutter, H. (2000). Learning and protection of proprietary assets in strategic alliances: Building relational capital. *Strategic Management Journal*, *21*(3), 217–237. https://doi.org/10.1002/(SICI)1097-0266(200003)21:3<217::AID-SMJ95>3.0.CO;2-Y

Kenny, D. A. & Milan, S. (2012). Identification: A nontechnical discussion of a technical issue. In R. H. Hoyle (ed.), *Handbook of structural equation modeling* (pp. 145–163). Guilford Press.

Kline, R. B. (2023). *Principles and Practice of Structural Equation Modeling*, 5th ed. Guilford Press.

Kühnel, S. (2001). The didactical power of structural equation modeling. In R. Cudeck, S. du Toit, & D. Sörbom (eds.), *Structural Equation Modeling: Present and Future. A Festschrift in Honor of Karl Jöreskog* (pp. 79–96). Scientific Software International.

Lang, K. M. & Little, T. D. (2018). Principled missing data treatments. *Prevention Science*, *19*(3), 284–294. https://doi.org/10.1007/s11121-016-0644-5

Little, T. D. (2013). *Longitudinal Structural Equation Modeling*. Guilford Press.

McDonald, R. P. & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*(1), 64–82. https://doi.org/10.1037/1082-989X.7.1.64

Mulaik, S. A. (2009). *Linear Causal Modeling with Structural Equations*. CRC Press.

Müthen, L. K. & Müthen, B. O. (1998–2017). *Mplus User's Guide*, 8th ed. Muthén & Muthén.

Boker, S., Nerale, M., Maes, H., et al. (2023). OpenMx: The OpenMx statistical modeling package. (R package 2.20.7). Available at. https://CRAN.R-project.org/package=OpenMx.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellora, D. T. (2018). The preregistration revolution. *PNAS*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Pek, J. & Hoyle, R. H. (2016). On the (in)validity of tests of simple mediation: Threats and solutions. *Social and Personality Psychology Compass*, *10*(3), 150–163. https://doi.org/10.1111/spc3.12237

Raykov, T. & Marcoulides, G. A. (2006). *A First Course in Structural Equation Modeling*, 2nd ed. Erlbaum.

Recio, L. A., Martín, P., Carvajal, F., Ruiz, M., & Serrano, J. M. (2013). A holistic analysis of relationships between executive function and memory in Parkinson's disease. *Journal of Clinical and Experimental Neuropsychology*, *35*(2), 147–159. http://dx.doi.org/10.1080/13803395.2012.758240

Rigdon, E. E. (2012). Rethinking partial least squares path modeling: In praise of simple methods. *Long Range Planning*, *45*(5–6), 341–358. https://doi.org/10.1016/j.lrp.2012.09.010

Rosseel, Y., Jorgensen, T. D., & Rockwood, N. (2022). lavaan: Latent variable analysis (R package 0.6-11). Available at: https://CRAN.R-project.org/package=lavaan.

SAS Institute Inc. (2021). *SAS/STAT 15.2 User's Guide*. SAS Institute Inc.

Sauvé, G., Kline, R. B., Shah, J. L., et al. (2019). Cognitive capacity similarly predicts insight into symptoms in first- and multiple-episode psychosis. *Schizophrenia Research*, *206*, 236–243. https://doi.org/10.1016/j.schres.2018.11.013

Shah, R. & Goldstein, S. M. (2006). Use of structural equation modeling in operations management research: Looking back and forward. *Journal of Operations Management*, ***24***(2), 148–169. https://doi.org/10.1016/j.jom.2005.05.001

StataCorp LLC (1985–2021). *Stata Structural Equation Modeling: Release 17*. Stata Press.

Steiger, J. H. (2001). Driving fast in reverse: The relationship between software development, theory, and education in structural equation modeling. *Journal of the American Statistical Association*, *96*(453), 331–338. https://doi.org/10.1198/016214501750332893

Tabachnick, B. G. & Fidell, L. S. (2013). *Using Multivariate Statistics*, 6th ed. Pearson.

Tarka, P. (2018). An overview of structural equation modeling: Its beginnings, historical development, usefulness and controversies in the social sciences. *Quality & Quantity*, *51*(1), 313–354. https://doi.org/10.1007/s11135-017-0469-8

Teo, T. (2010). A case for using structural equation modelling (SEM) in educational technology research. *British Journal of Educational Technology*, *41*(5), 89–91. https://doi.org/10.1111/j.1467-8535.2009.00999.x

Textor, J., van der Zander, B., & Ankan, A. (2020). dagitty: Graphical analysis of structural causal models (R package 0.3-0.). Available at: https://CRAN.R-project.org/package=dagitty.

Thelwall, M. & Wilson, P. (2016). Does research with statistics have more impact? The citation rank advantage of structural equation modeling. *Journal of the Association for Information Science and Technology*, *67*, 1233–1244. https://doi.org/10.1002/asi.23474

Tomarken, A. J. & Waller, N. G. (2003). Potential problems with "well-fitting" models. *Journal of Abnormal Psychology*, *112*(4), 578–598. https://doi.org/10.1037/0021-843X.112.4.578

Westfall, P. H., Henning, K. S. S., & Howell, R. D. (2012). The effect of error correlation on interfactor correlation in psychometric measurement. *Structural Equation Modeling*, *19*(1), 99–117. http://dx.doi.org/10.1080/10705511.2012.634726

Williams, L. J. (2012). Equivalent models: Concepts, problems, alternatives. In R. H. Hoyle (ed.), *Handbook of Structural Equation Modeling* (pp. 247–260). Guilford Press.

Williams, T. C., Bach, C. C., Matthiesen, N. B., Henriksen, T. B., & Gagliardi, L. (2018). Directed acyclic graphs: A tool for causal studies in paediatrics. *Pediatric Research*, *84*(4), 487–493. https://doi.org/10.1038/s41390-018-0071-3

Wold, H. (1982). Soft modeling: The basic design and some extensions. In K. G. Jöreskog & H. Wold, (eds.), *Systems Under Indirect Observations: Part II* (pp. 1–54). North-Holland.

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety.

Educational and Psychological Measurement, *73*(6), 913–934. https://doi.org/10.1177/0013164413495237

Wright, S. (1920). The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs. *Proceedings of the National Academy of Sciences*, *6*(6), 320–332. https://doi.org/10.1073/pnas.6.6.320

von Oertzen, T., Brandmaier, A. M., & Tsang, S. (2015). Structural equation modeling with Ωnyx. *Structural Equation Modeling*, *22*(1), 148–161. https://doi.org/10.1080/10705511.2014.935842

Xia, Y. & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods*, 51(1),409–428. https://doi.org/10.3758/s13428-018-1055-2

Yamaga, E., Sato, Y., & Minakuchi, S. (2013). A structural equation model relating oral condition, denture quality, chewing ability, satisfaction, and oral health-related quality of life in complete denture wearers. *Journal of Dentistry*, *41*(8), 710–717. https://doi.org/10.1016/j.jdent.2013.05.015

Zhang, M. F., Dawson, J., & Kline, R. B. (2021). Evaluating the use of covariance-based structural equation modelling with reflective measurement in organisational and management research: A review and recommendations for best practice. *British Journal of Management*, *32*(2), 257–272. https://doi.org/10.1111/1467-8551.12415

25 Structural Equation Modeling

Rex B. Kline

Abstract

Structural equation modeling (SEM) is a family of statistical techniques and methods for testing hypotheses about causal effects among observed or proxies for latent variables. There are increasing numbers of SEM studies published in the research literatures of various disciplines, including psychology, education, medicine, management, and ecology, among others. Core types of structural equation models are described, and examples of causal hypotheses that can be tested in SEM are considered. Requirements for reporting the results of SEM analyses and common pitfalls to avoid are reviewed. Finally, an example of evaluating model fit is presented along with computer syntax so that readers can reproduce the results.

Keywords: Structural Equation Modeling; Covariance Structure Analysis; Covariance-Based SEM; Causal Models

Introduction

Structural equation modeling (SEM) is a family of multivariate statistical techniques for estimating presumed causal relations in either observational or experimental studies. The terms path analysis, confirmatory factor analysis, and latent growth curve modeling, among others, all refer to particular types of SEM analyses. It combines aspects of (1) factor analysis, which estimates latent variables (theoretical concepts), given data from their indicators, or observed (manifest) variables (see Chapter 24 in this volume); and (2) regression analysis, that analyzes multiple explanatory variables of the same response (outcome) variable while controlling for intercorrelations among all variables. Equations for multiple variables are simultaneously analyzed such that an outcome variable in one equation can be specified as causal variable in a different equation. Variables can be either observed or latent, and the distinction between observed and latent variables can take account of measurement error. It is also possible to analyze means in SEM, including the comparison of means from independent samples or dependent samples (e.g., repeated measures) on observed variables or proxies for latent variables (Bagozzi & Yi, 2012).

The SEM family is also flexible in that it accommodates analyses that are more exploratory. It can be used in *model generation*, where an initial model is found to be inconsistent with the data and is subsequently modified over a series of follow-up analyses. A second context is testing *alternative models*, where two or more a priori models, comprised of the same variables but specified based on different theories, are all fitted to the same data. A third context is *strictly confirmatory* – a single model that is either retained or rejected based on its correspondence with the data with no further analysis (Jöreskog, 1993). In any context, the goals are to (1) understand patterns of covariances among a set of measured variables, and (2) explain as much of their variance as possible with a statistical model that makes theoretical sense, is parsimonious, and has acceptably close correspondence to the data.

The combination of features just described is relatively unique and probably explains why SEM is being applied in rapidly increasing numbers of studies in disciplines that include psychology, education, medicine, ecology, environmental sciences, commerce, marketing, international business, management, operations research, tourism, and sustainable manufacturing, among others (e.g., Teo, 2010; Thelwall & Wilson, 2016). Many graduate programs now offer courses in SEM, and there are numerous summer schools and seminars on SEM for established researchers around the world. The growing availability of computer tools, some free of charge, has also made SEM more accessible to applied researchers. With no exaggeration, it can be said that SEM is becoming an essential set of statistical techniques.

There are also downsides to the increasing use and popularity of SEM – students or established researchers may be pressured by supervisors or reviewers to use SEM as a cutting-edge method when a simpler statistical technique would do. A related concern is that SEM could be used with little understanding, especially if an emphasis on ease of use of statistical computer tools gives beginners the false impression that SEM is easy (Steiger, 2001). Another problem in many articles is that so much attention is paid to technical aspects of applying SEM that the theoretical sense and meaning of the hypotheses behind the model are neglected. The risk is that, as Tarka (2018, p. 342) put it, "the use of SEM . . . is full of overuse, incorrect interpretation and overinterpretation" – due to apparent unawareness of its potential limitations.

There is also evidence for widespread deficient reporting of results from SEM analyses (e.g., Fan et al., 2016; Shah & Goldstein, 2006). For example, some colleagues and I reviewed a total 144 SEM studies published in 12 top organizational and management journals from 2011 to 2016 (Zhang et al., 2021). Each article was evaluated against criteria that included the clarity of the rationale for using SEM versus alternative techniques, whether hypotheses were tested in a clear and specific order, and whether statistical results were described in sufficient detail. Many shortcomings were apparent: explicit justification for using SEM instead of alternative methods was given in about 40% of the studies; data screening or distributional assumptions were explicitly described in about 20%; and complete details about model fit to the data were reported in about 20% of reviewed studies. That is, the reader of the typical study was not provided with enough information to understand whether the findings actually had any meaningful interpretation. As a reviewer of SEM manuscripts for about 30 different journals, I see these reporting problems all the time.

SEM Families

The term “SEM” does not refer to a single set of statistical techniques or methods. Instead, the three distinct bodies of work listed next, and described afterward, make up what we now refer to as SEM:

1. *Covariance-based SEM* (CB-SEM), also called *covariate structure analysis* or *covariance structure modeling*, is the form of SEM most familiar in psychology, sociology, education, and related disciplines.
2. *Variance-based SEM* (VB-SEM), also known as *composite SEM* or *partial least squares path modeling* (PLS-PM), among other names, is more familiar in research areas on management, organization, and marketing.
3. The *structural causal model* (SCM) originated in Pearl’s (2009) work on Bayesian networks in the 1980–1990s and since extended to the more general problem of causal modeling. The SCM is probably best known in epidemiology, health sciences, and computer science, but that is changing.

All three statistical families just listed owe their origins to pioneering work by the geneticist Sewall Wright (1920), who developed the technique of *path analysis* for testing hypotheses about causal effects among a set of variables. Diagrams for Wright’s models included both observed and latent variables, and they are remarkably similar to modern path diagrams (e.g., Wright, 1920, p. 328). Wright’s work on estimating causal effects using regression methods was introduced to the social and behavioral sciences in the 1960 and 1970s (Tarka, 2018), and the very first computer program for SEM available on mainframe computers, LISREL III (Jöreskog & Sörbom, 1976), combined regression and factor analysis methods, both of which analyze covariances (hence the term “CB-SEM”). All modern SEM computer tools – including the most recent version of LISREL itself (Jöreskog & Sörbom, 2021) – share LISREL III as a forerunner.

The VB-SEM family dates to an approach called “soft modeling,” developed in the 1970s and 1980s by Wold (1982), which estimates latent variables as weighted linear combinations of observed variables, or composites (also called components). Statistical methods for composites are generally less demanding than those that approximate latent variables as common factors (i.e., CB-SEM). For example, composite-based methods may require smaller samples compared with CB-SEM techniques. Analyses of composite models are also less prone to technical problems, such as the failure of iterative estimation to reach a stable solution. Analyses in VB-SEM maximize prediction (i.e., R^2) of outcome variables. In contrast, CB-SEM methods aim to maximize the overall correspondence between model and data; this may not necessarily maximize prediction for individual outcomes. If maximizing prediction is a primary goal, then VB-SEM may be preferred over CB-SEM (Rigdon, 2012).

Pearl’s (2009) SCM corresponds to *non-parametric SEM,* where causal hypotheses are represented in a *directed acyclic graph* (DAG). A DAG is non-parametric because it assumes no particular operationalization for any theoretical variable or any specific functional form of statistical association between variables (e.g., linear versus curvilinear relations for continuous variables). Instead, there are methods in the SCM to analyze a DAG to determine whether it is possible to estimate a target causal effect through the inclusion of other variables as covariates or as instrumental variables, among other possibilities. Thus, a DAG is not a static entity in the SCM; instead, it can be analyzed *with no data whatsoever*, and insights from analyzing the graph can be invaluable in dealing with potential confounding (see Williams et al., 2018 for examples in pediatrics).

It is beyond the scope of this chapter to cover all three SEM families. Instead, just CB-SEM is considered from this point, for two reasons: (1) I would wager that more researchers are familiar with CB-SEM than with VB-SEM or the SCM (Thelwall & Wilson, 2016), and (2) knowing the concepts of CB-SEM provides a strong basis for learning about the other two approaches (Astrachan et al., 2014). This is because knowing something about CB-SEM means that the researcher must (1) understand how concepts are defined, operationalized, and expressed in scores from imperfect observed measures; and (2) comprehend basic principles of regression analysis, factor analysis, and the correct interpretation of standard errors and statistical significance (Kline, 2023; Kühnel, 2001). Ideally, researchers who work with colleagues in other disciplines should know about all three families, but CB-SEM is a good place to start. In the rest of the chapter, the term "SEM" refers to CB-SEM.

SEM Steps and Reporting Standards

Described next are the six basic steps in SEM. They are actually iterative because problems at a particular step may require a return to an earlier step. The context of model generation is assumed.

Step 1: Specification

Specification is the representation of the researcher's hypotheses as a series of equations or as a model diagram (or both). This involves defining the observed and latent variables and their presumed relations. Outcome (dependent) variables in SEM are referred to as *endogenous variables*. Every endogenous variable has at least one presumed cause among other variables in the model, and error terms that represent unexplained variation are typically associated with each endogenous variable. Depending on the hypothesis, an endogenous variable could be specified as a cause of a different endogenous variable. Endogenous variables, as just described, are *intervening variables* – they are specified as affected by causally prior variables, and in turn they affect other variables further "downstream" in a causal pathway. In contrast, exogenous variables are strictly causal – whatever causes them is not represented in the model.

Whether a variable is endogenous or exogenous is determined solely by the theory being tested. This means that the model is specified *before* the data are collected, and the whole model represents the total set of hypotheses to be evaluated in the analysis. *Specification is the most important step*. This is true because results from the analysis assume that the model is correct. Because the initial model is not always retained, I suggest that, before data collection, researchers make a list of possible modifications that would be *justified according to theory*; that is, prioritize the hypotheses, represent just the very most important ones in the model, and leave the rest for a "backup list." Preregistration of the analysis plan would make strong a statement that changes to the initial model were not made after the examining the data (Nosek et al., 2018).

Step 2: Identification

Identification concerns the issue of whether each model parameter can be expressed as a unique function of the variances, covariances, or means in a *hypothetical* data matrix. It is basically a mathematical proof expressed in *symbolic* form that there is potentially a unique estimator for each effect in the model. *Thus, identification has nothing to do with real data (numbers) or with sample size*. Instead, it is an inherent property of the model, given all the equations that define it. A model that is not identified remains so regardless of both the data matrix and the sample size ($N = 100$, 1,000, etc.) and, thus, must be respecified. An intuitive example follows.

Consider these formulas:

$$\begin{aligned} a + b &= 6 \\ 3a + 3b &= 18 \end{aligned} \tag{25.1}$$

Equation 25.1 is not identified because there is no unique solution (for a, b) that satisfies both formulas. Instead, there are *infinite* solutions, such as (4, 2), (5, 1), and so on. This happens because the second formula in Equation 25.1 is linearly dependent on the first formula – an inherent characteristic. Now consider the formulas listed next where the second is not linearly dependent on the first:

$$\begin{aligned} a + b &= 6 \\ 2a + b &= 10 \end{aligned} \tag{25.2}$$

Equation 25.2 has a single solution – it is (4, 2) – so the whole expression is identified. Structural equation models are typically more complex than Equations 25.1 and 25.2 (i.e., it is often impractical to inspect individual parameters especially in large models). Instead, there are graphical methods and identification heuristics that can determine whether some, but not all, models are identified (Kenny & Milan, 2012). There are also computer tools that analyze diagrams of path models for identification in the SCM approach to SEM (Textor et al., 2020).

Step 3: Measure Selection and Data Collection

Measure selection and data collection are essential activities in most empirical studies. See Kline (2023, ch. 4) and Lang and Little (2018) for how to select measures and deal with potential data-related problems in SEM (e.g., missing data, univariate or multivariate outliers, and extreme collinearity).

Although there have been efforts to make the application of SEM in smaller samples more feasible (Deng et al., 2018), the reality is that SEM is a large-sample technique. Unfortunately, there is no simple answer to the question of how large a sample is needed. This is because sample size requirements vary with model size or type, estimation method, distributional assumptions, and level of measurement for outcome variables, among other considerations. For example, estimation methods in SEM with no distributional assumptions generally need larger samples than methods that assume normal distributions. Larger models with more variables and effects may require larger samples than smaller, simpler models. There is also evidence that sample sizes in many, if not most, published SEM studies are too small in terms of both precision and statistical power (Wolf et al., 2013). As a rule of thumb, $N = 200$ or so might be a reasonable minimum sample size for smaller, more basic models (Barrett, 2007), but 200 is not a magic number. The requirement for large samples complicates replication in SEM, especially when studying rare populations, such as patients with a low-base-rate illness. In this case, it could be challenging to collect a sufficiently large sample for a single analysis, much less twice the number of cases for an additional cross-validation sample, where a model is analyzed in the original sample, and then these analyses are replicated in the second sample of equal size.

Step 4: Analysis

Analysis is carried out using an SEM computer program to fit the model to the data. A few things take place at this step. First, (a) evaluate model fit to determine how well the model fits the data. Often, the initial model does not adequately explain the data; if so, skip the rest of this step and go to step 5. Otherwise, next (b) interpret the parameter estimates, and (c) consider *equivalent models* that fit the data *exactly* as well as the researcher's model but feature contradictory hypotheses about causation among the same variables (Henley et al., 2006). An example is presented later, but the failure to acknowledge equivalent models is a widespread problem in SEM studies that is also a form of confirmation bias.

Step 5: Respecification

In this step, the initial model is altered and fitted to the same data, *but any respecified model must be theoretically justified* (i.e., consult the backup list mentioned earlier). If there is no such justification, it may be better – and more honest, too – to retain no model (Hayduk, 2014), especially compared with making changes solely to improve the fit of the model in a particular sample. The problem is that post doc, data-driven respecification can lead to a model that does not replicate because it capitalizes so strongly on sample-specific variation.

Step 6: Reporting

This is the written summary of the results. If a model is retained, describe both global fit and local fit. *Global fit* concerns the overall or average match between the model and the data matrix. Just as averages do not indicate variability, models in SEM with apparently satisfactory global fit can have problematic *local fit*; this is measured by residuals calculated for every pair of measured variables (see Tomarken & Waller, 2003 for examples). Residuals in SEM concern differences between observed (i.e., in the data) versus predicted (i.e., from the model) covariances or correlations, and as absolute residuals increase in size, local fit becomes worse. The analogy in regression is the difference between R^2 – overall predictive power (global fit) – and regression residuals – differences between observed and predicted scores. Aberrant patterns of regression residuals indicate a problem in the analysis even if the value of R^2 is reasonably high. Just as reports about regression results with no mention of the residuals are incomplete, so too are reports in SEM in which only global model fit is described. For an example of full reporting on residuals in SEM, see Sauvé et al. (2019, Appendix A).

Reporting about both global and local model fit is part of journal article reporting standards for SEM studies by the American Psychological Association (Appelbaum et al., 2018) and this is based on earlier standards for SEM by Hoyle and Isherwood (2013) for the journal *Archives of Scientific Psychology*. Reporting standards also call on researchers to

(a) outline how the sample size was determined, such as through power analysis
(b) give a full account of model specification, including the rationale for hypotheses about the directionality of causal effects (i.e., X causes Y and not the reverse), all in the context of relevant theory
(c) explain the bases for respecification of an initial model and whether respecifications were a priori or post hoc
(d) interpret statistical results according to evidence-based criteria
(e) justify the preference for any retained model over equivalent models that explain the data just as well
(f) report the unstandardized solution with standard errors and the standardized solution
(g) report sufficient summary statistics to allow secondary analysis or make the raw data file available.

SEM Computer Programs

In the late 1970s, LISREL was among a small number of computer tools for SEM, but today there are many options for SEM software, both commercial and freely available. Free software packages for SEM include lavaan (Rosseel et al., 2022) and OpenMx (Boker et al., 2022) for the R computing environment. There are also R packages for conducting specialized types of SEM analyses, such as semTools for simulation and power analysis (Jorgensen et al., 2022). Other free options that do not involve R include JASP, an integrated, open-source application with capabilities for traditional (frequentist) and Bayesian analyses (including SEM; JASP Team, 2022), and Ωnyx (pronounced “onyx”), which features a drawing editor where the user specifies the model and controls the analysis by drawing the model on the computer screen (von Oertzen et al., 2015).

Free-standing commercial products for SEM analyses include Amos, EQS, Mplus, and LISREL (respectively, Arbuckle, 2021; Bentler & Wu, 2020; Müthen & Müthen, 1998–2017; Jöreskog & Sörbom, 2021). Some widely used software for general statistical analyses have procedures, functions, or commands for SEM. Examples include the sem command in Stata (StataCorp, 1985–2021) and the CALIS procedure in SAS/STAT (SAS Institute, 2021). Some universities and research centers have site licenses for commercial SEM software that allow free use by researchers and students, but individual licenses can be relatively expensive. Commercial products have the advantage of complete manuals with many analysis examples or data sets, if cost is no problem; otherwise, free SEM software (e.g., lavaan) is nearly as capable as commercial products.

Core Types of Models

Described next are three core types of models in SEM with examples of each from actual studies. All example models are identified. Presented in Figure 25.1 is the *manifest-variable (classical) path model* analyzed by Yamaga et al. (2013). Such models feature *single-indicator measurement*, where each construct is measured by a single observed variable. For all examples,

(1) observed variables are represented with squares or rectangles
(2) latent variables are depicted with circles or ovals
(3) lines with single arrowheads point from presumed causes to endogenous variables
(4) presumed covariances between measured variables are represented as curved lines with arrowheads at each end.

It is also common in model diagrams to represent error terms for endogenous variables, but there is no standard symbolism for doing so. Perhaps the most basic symbol is a line with a single arrowhead oriented at a 45-degree angle that points to each outcome (e.g., ↙; see Figure 25.1), but McDonald and Ho (2002) describe additional ways to graphically represent error terms in diagrams of structural equation models.

In a sample of 166 edentulous (toothless) dental patients who presented themselves for complete denture therapy, Yamaga et al. (2013) measured the integrity of mandibular ridge form (lower jaw bone formation), retention and stability of mandibular complete denture, jaw relation (whether the cusps of opposing teeth on the lower and upper [maxillary] jaws correctly interlock), perceived chewing ability (mastication), satisfaction with exiting complete dentures, and extent of oral health problems. Their path model in Figure 25.1 represents the hypotheses that

(1) ridge form, retention, and stability all co-vary and also directly affect jaw relation
(2) jaw relation, in turn, is a direct cause of both mastication and denture satisfaction
(3) mastication is also caused by stability, and satisfaction is also affected by both ridge form and mastication
(4) oral health problems are directly affected by both mastication and satisfaction.

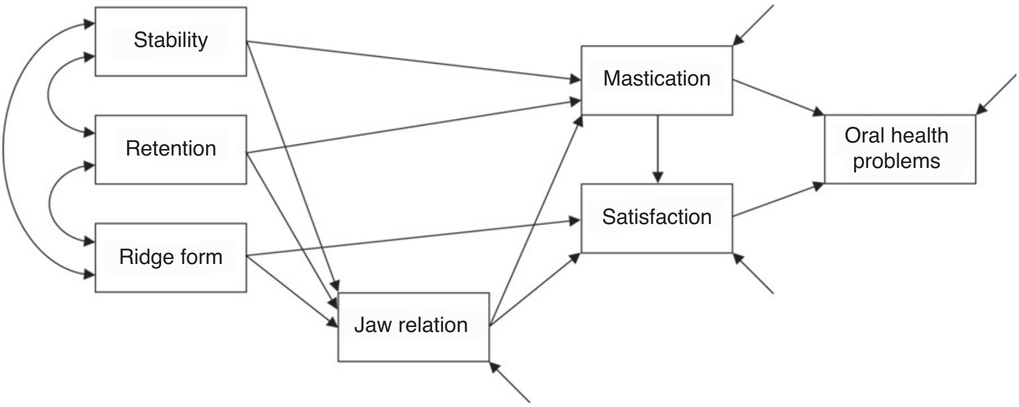


Figure 25.1 *Example of a manifest-variable path model analyzed by Yamaga et al. (2013).*

The variables jaw relation, mastication, and denture satisfaction in Figure 25.1 are specified as intervening variables that "absorb" effects from prior causal variables and "transmit" those effects to subsequent outcomes. For example, the indirect pathway

stability → jaw relation → mastication

represents the hypothesis that stability of mandibular complete denture affects jaw relation, which, in turn, impacts mastication. Indirect effects are part of the concept of mediation, *but the two are not synonymous*. This is because *mediation* is the strong causal hypothesis that one variable (stability) causes *changes* in another variable (jaw relation), which leads to *changes* in an outcome (mastication; Little, 2013). The emphasis on "changes," in the definition just stated, highlights the requirement for *time precedence* – measurement of presumed causes before their outcomes. With no time precedence, it is difficult to interpret estimates for indirect effects as evidence for mediation (Pek & Hoyle, 2016). Yamaga et al.'s (2013) design was cross-sectional – all variables were measured at the same occasion (see Chapter 13 in this volume) – so the term "mediation" does not automatically apply to any of the indirect causal pathways in Figure 25.1.

Especially in cross-sectional designs, which have no inherent support for causal inference, *directionalities of causal effects in SEM are assumed, not tested*. This is because there is little, if anything, from analysis that could either disconfirm or verify hypotheses about causal priority. For example, outcomes of significance testing for the *path coefficient* of *X* → *Y*, a presumed direct effect of *X* on *Y*, could fail to be significant in a small sample due to insufficient power. The phenomenon of equivalent models with the opposite specification – *Y* → *X* – which fit the data just as well as the original model, discounts the possibility that significant path coefficients prove causation. This is why it is critical to provide clear and reasoned justifications for directionality specifications, especially in cross-sectional designs. Thus, SEM is not a technique for causal discovery. This means that, if given a true model, SEM could be applied to estimate the magnitudes of causal effects represented in the model. However, this is not how SEM is typically used; instead, a causal model is *hypothesized*, and the model is fitted to sample data *assuming* that all its specifications are correct.

Other assumptions of the path model in Figure 25.1 are briefly summarized next: (1) Score reliabilities on the exogenous variables (stability, retention, and ridge form) are perfect – $r_{XX} = 1.0$. Exogenous variables in path models, as in the figure, have no error terms, so there is no "room" for measurement error in these variables. This requirement does not apply to endogenous variables (e.g., jaw relation) that have error terms that absorb measurement error. (2) There are no unmeasured common causes, or confounders, for any pair of variables in the model. This assumption is required because the omission of confounders can seriously bias values of coefficients in both regression analysis and SEM (Cohen et al., 2003). (3) The error terms in Figure 25.1 are independent; that implies all unmeasured causes of endogenous variables are all pairwise uncorrelated and also with all three exogenous variables. Altogether, these assumptions are very demanding. Results from analysis of the model in Figure 25.1 are described in the last section of this chapter.

Figure 25.2 shows a *confirmatory factor analysis* (CFA) model analyzed by Filippetti and Krumm (2020), who administered five performance tasks, hypothesized to reflect two dimensions of cognitive flexibility, to 112 children aged 8–12 years. These domains included reactive flexibility – the capability to modify behavior – and spontaneous flexibility– the ability to generate novel responses. The two language-based fluency tasks in the figure involve asking examinees to say as many words as possible for two categories (e.g., animals; semantic fluency) or starting with a specific letter (e.g., S; phonetic fluency) for 60 seconds. The pattern fluency task measures the ability to produce unique geometric designs within a time limit. The three tasks just described are specified as indicators of spontaneous flexibility; this is represented in Figure 25.2 with the symbol for a latent variable – an oval (circles can also designate latent variables in model diagrams). Indicators in CFA models have error terms that capture random measurement error in the observed variables. Thus, it is *not* assumed in CFA that the scores are perfectly precise.

The remaining two observed variables in Figure 25.2 are specified as indicators of reactive flexibility. These tasks include a computerized card-sorting task, where examinees are asked to match geometric patterns. Because they are told only whether their responses are correct or incorrect, examinees must infer the matching rules. The trail-making task requires examinees to draw lines in an alternating series of numbers and letters in sequential order (e.g., 1, A, 2, B, and so on). Numerals (e.g., 1) that appear in the figure next to certain direct effects (one per factor) are *scaling constants* that specify metrics for the factors. For example, the specification in the figure

$$\text{reactive flexibility} \rightarrow \text{card sorting} = 1$$

assigns a scale to the reactive flexibility factor. That scale corresponds to variation in the card-sorting task that is explained by the factor it is presumed to measure. It is

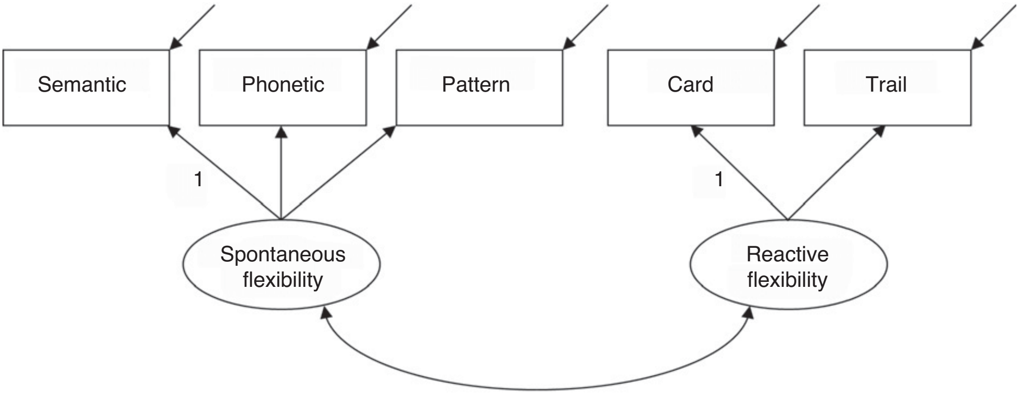


Figure 25.2 *Example of a CFA model analyzed by Filippetti and Krumm (2020).*

usually arbitrary which direct effect is so specified, but latent variables must be scaled before the computer can derive statistical estimates about them (see Brown, 2015, for discussion of other options to scale factors).

The symbol for a covariance that connects the spontaneous flexibility and reactive flexibility factors in Figure 25.2 instructs the computer to estimate their covariance (unstandardized) or correlation (standardized), given the model and data. It is often reasonable to assume that hypothetical constructs are related, such as cognitive ability factors (e.g., verbal, visual–spatial, memory), and covariances between all pairs of factors are routinely estimated in CFA. If two factors are believed to be independent, their covariance can be specified as zero to test this hypothesis (see Brown, 2015 for examples). In addition to the two-factor, five-indicator structure represented in Figure 25.2, the model also assumes that (1) omitted causes of the indicators are unrelated to the factors, and (2) omitted causes for each indicators have no overlap with those for all other indicators.

It is important, in any method of factor analysis, to avoid the *naming fallacy*; just because a factor is named does not mean that the corresponding hypothetical construct is understood or even correctly labeled. For instance, the label “reactive flexibility” in Figure 25.2 does not preclude other interpretations of what the card-sorting and trail-making tasks measure (e.g., abstract reasoning or visual analysis). Factor labels are conveniences that are more “reader friendly” than abstract symbols, but they are not substitutes for critical thinking (Kline, 2023). Another potential error is *reification* – the false belief that a factor *must* correspond to something in the real world. Factors are statistical abstractions from observed measures, and whether such abstractions describe any tangible entity, dimension, or process is an open question (Rigdon, 2012).

Figure 25.3 shows a *structural regression (SR) model*, also called a *latent-variable path model* or *full-LISREL model* because LISREL was one of the first computer programs to analyze such models. The SR model in the figure was analyzed by Recio et al. (2013), who administered measures of executive function – cognitive processes needed for monitoring and control of behavior (e.g., attentional focus) – and

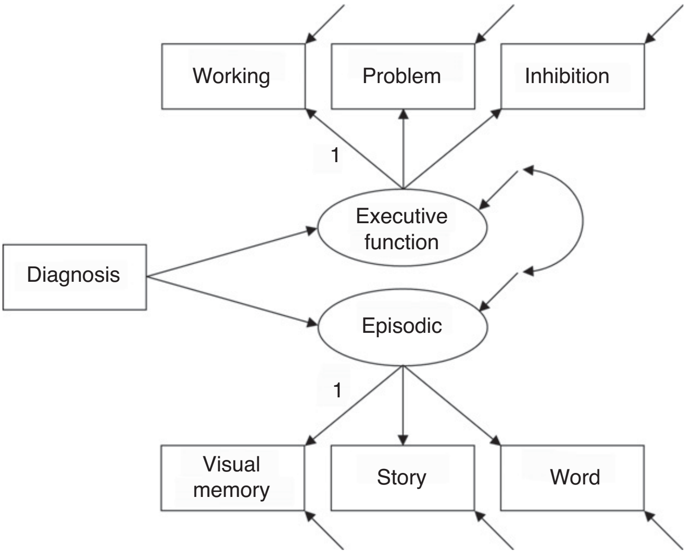


Figure 25.3 *Example of a SR model analyzed by Recio et al. (2013).*

measures of episodic memory – recall of visual or auditory stimuli – within samples of patients with Parkinson's disease and neurologically healthy adults matched for age and level of education. It is worth noting that the group sizes were, I believe, too small – a total of 23 patients and 18 control cases – for precise estimation of the model in Figure 25.3. A more reasonable group size would be $n \geq 100$, but even that number may be inadequate for sufficient statistical power.

The measurement part of the model corresponds to the two factors, each with three indicators: working memory, problem solving, and inhibition for the executive function factor, and tests of visual, story, and word recall for the episodic memory factor. Both factors just mentioned are specified as outcomes of the dichotomous variable of diagnosis, which specifies membership in either the Parkinson's disease group or the control group. Because the factors are endogenous in Figure 25.3, they each have error terms that represent variation not explained by diagnosis. In contrast, factors in CFA models are exogenous and do not have error terms (cf. Figure 25.2).

The curved line with arrowheads at each end in Figure 25.3 represents an *error covariance* in the unstandardized solution or an *error correlation* in the standardized solution. This specification instructs the computer to estimate the association between the executive function and episodic memory factors *after controlling for diagnosis*. Here it makes sense that the two cognitive factors would be related above and beyond the distinction between Parkinson's disease and control cases. Other valid reasons to specify correlated error terms in SEM include autocorrelation among variables in longitudinal designs, common response sets (systematic difference in how participants respond to questions regardless of item content), and shared stimuli over tasks (Westfall et al., 2012). Each error correlation added to a model makes it more complex and generally improves fit. A concern is that error correlations are added mainly to enhance fit without substantive reasons. As with any other model specification, the inclusion of correlated errors requires justification.

Example SEM Analysis and Reporting Recommendations

In their original analysis of the path model in Figure 25.1, Yamaga et al. (2013, p. 14) reported sufficient summary statistics for their raw data – correlations and standard deviations – to allow other researchers to reproduce their results in a secondary analysis (with slight rounding errors). Doing so is a best practice both in SEM and other types of quantitative studies (Appelbaum et al., 2018). This is because, even with no access to the raw data, other researchers can independently verify the original analyses or test hypotheses not considered by the authors of the original work. There are some types of SEM analyses that require raw data files. Examples include the analysis of continuous variables with methods that adjust for severely non-normal distributions or the analysis of ordinal data (see Kline, 2023, for more information), but summary statistics are all that's needed in this example. Listed in the appendix at the end of this chapter is syntax for lavaan that fits the model in Figure 25.1 to summary statistics reported by Yamaga et al. (2013) for *N* = 166 cases. This syntax can be executed in R after installing the lavaan package – install.packages("lavaan", dependencies = TRUE). The output file will contain all the results described next. The estimation method is default maximum likelihood and assumes normal distributions.

Listed next is a suggested structure for the results section that is also consistent with reporting standards for SEM (Appelbaum et al., 2018). It is assumed that the theoretical rationale for model specification is outlined earlier in the manuscript:

(1) Explicitly tabulate numbers of observations, free model parameters, and model degrees of freedom.
(2) Report results about both global model fit and local model fit, or the residuals.
(3) Justify the decision to either retain the model as initially specified, reject the model before the analysis enters a respecification phase, or reject the model with no further changes nor analyses. If a respecified model is retained, state the rationale for any modifications to the initial model, including whether respecification was mainly a priori or empirical.
(4) If a model is retained, then (a) report the unstandardized parameter estimates with standard errors and the standardized solution. Also, (b) directly acknowledge the existence of equivalent models, generate at least a few examples, and argue why the retained model is preferable to any equivalent version with exactly the same fit to the data.

Model Degrees of Freedom

The *model degrees of freedom*, df_M, is the difference between the number of observations and the number of free model parameters. The number of *observations* for continuous variables, when means are not analyzed (as in this example), equals v (v + 1)/2, where v is the number of observed variables. For example, $v = 7$ in Figure 25.1, so the number of observations is 7(8)/2, or 28; this equals the number of elements in the covariance matrix generated by the descriptive statistics in Yamaga et al. (2013, p. 14) in lower diagonal form, where redundant values above the diagonal are eliminated (see the appendix).

A *free parameter* is estimated by the computer with the sample data. Free parameters when means are not analyzed include (1) variances and covariances of exogenous variables, (2) direct effects on endogenous variables from other variables in the model (but not error terms), and (3) the variance of each error term (Kline, 2023). In Figure 25.1, there are three exogenous variables with a covariance between each pair, so the total number of variances and covariances here is 3 + 3 = 6. There are four endogenous variables in the figure with a total of 11 direct effects on them from other variables. Each endogenous variable has an error term; a total of four error variances must be estimated by the computer. Thus, the total number of free parameters is

$$6 + 11 + 4 = 21 \text{ so } df_M = 28 - 21 = 7.$$

Models with no degrees of freedom ($df_M = 0$) will perfectly fit the data. This is because such models are as complex, in terms of free parameters versus observations, as the data they are supposed to explain. Models where $df_M = 0$ test no particular hypothesis and, thus, are rarely of interest. Positive degrees of freedom ($df_M > 0$) allow for the *possibility* of discrepancies between model and data – imperfect fit. A key question in the analysis for models with $df_M > 0$ is whether expected differences between model and data are so great that the model should be rejected. Thus, $df_M > 0$ is an effective requirement in SEM, and there is a preference for models with greater degrees of freedom or models that are more parsimonious based on df_M (Raykov & Marcoulides, 2006). Models with negative degrees of freedom ($df_M < 0$) are not identified and must be respecified so that $df_M \geq 0$ before they can be analyzed.

Global Fit

There are two kinds of global fit statistics in SEM: model test statistics (i.e., significance tests) and approximate fit indexes; these are not significance tests. The most widely reported test statistic is the *model chi-square* with its degrees of freedom, df_M. The statistic is designated here as chi_M. It tests the null hypothesis that the researcher's model perfectly fits the *population* data matrix. The value of chi_M equals the product of sample size (N) and the degree of difference between the sample data matrix and associations for the same variables predicted by the researcher's model. If $\text{chi}_M = 0$, the model perfectly fits the sample data matrix. As model–data discrepancies increase, the value of chi_M increases, too.

If $\mathrm{chi}_M > 0$ and its p-value is less than α, the criterion level of statistical significance, then (1) the null hypothesis of perfect fit is rejected, and (2) the model *fails* the chi-square test. Suppose that $\mathrm{chi}_M = 12.50$ for a model where $df_M = 5$. The p-value for this result is 0.029. If $\alpha = 0.05$, the model fails the chi-square test because $p < \alpha$. This means that the difference between the data matrix and the predicted matrix is significant at the 0.05 level. *Passing* the model chi-square test happens whenever $p \geq \alpha$, such as chi_M (5) = 10.50, $p = 0.062$ (when testing at the 0.05 level). *But passing the chi-square test does not automatically mean that the model also has satisfactory local fit*. It can and does happen, especially in samples that are not large and where the power of the chi-square to detect appreciable model–data discrepancies is low, that passing models have poor local fit; that is, the residuals are problematic. *Such models should not be retained even though they passed the chi-square test*. Likewise, it can happen, in very large samples, that a model fails the chi-square test, but the residuals indicate trivial discrepancies in local fit. In this case, the researcher might reasonably argue to retain the model, given satisfactory residuals. In fact, some researchers used to divide chi_M by df_M to reduce its sensitivity to sample size, but (1) df_M has nothing to do with sample size, and (2) there are never any specific values of chi_M/df_M that indicate "good" fit (e.g., < 3.0, 5.0, or some other value). Therefore, I do not recommend it.

A widespread but poor practice in published SEM studies occurs when (1) the model fails the chi-square test but (2) the researcher automatically dismisses this result because "the model chi-square is affected by sample size" or some such rationale that is actually false; N affects chi_M *only when the model is wrong* (Hayduk, 2014). Failing the chi-square test should be interpreted as indicating covariance evidence against the model, and that failure should be thoroughly diagnosed (i.e., inspect the residuals). Passing the chi-square test should also be followed by careful inspection of the residuals *because the details of fit are in the residuals*.

Approximate fit indexes are continuous measures of model–data discrepancy. Some approximate fit indexes are scaled like chi_M – a value of zero is the best result concerning model fit. Others have more-or-less standardized metrics from 0 to 1.0, where 1.0 is the best result. Dozens of approximate fit indexes have been described in the literature, and output for some SEM computer tools includes values for rather lengthy lists of such indexes. A problem with all approximate fit indexes is that there is little correspondence between their numerical values and types or seriousness of specification error (Hayduk, 2014). The same thing is true about chi_M, its p-values, and the residuals. One reason is equivalent models that have identical values of all global fit statistics *and* residuals even though they represent contradictory sets of hypotheses.

An issue with approximate fit indexes is overreliance on now-discredited fixed thresholds that supposedly indicate whether model–data correspondence is "good." An example of a "golden rule" for the hypothetical "ABC" global fit statistic is, "if ABC > 0.95, then model fit is good." Such thresholds date from computer simulation studies in the 1980s and 1990s about a very narrow range of models, but subsequent results indicated that these fixed thresholds do not always apply to other kinds of models or data (Barrett, 2007). For example, fixed thresholds were originally developed for models with continuous endogenous variables, but they are not accurate for models with ordinal endogenous variables (Xia & Yang, 2019). There is no problem with reporting values of approximate fit indexes, *but there are no magic cutting points that somehow differentiate between models with "good" versus "poor" fit*, especially if the researcher does not also look to the residuals for more detailed information about model fit.

Listed next is what I believe is a minimal set of approximate fit indexes that should be reported in most analyses (Kline, 2023, ch. 10). Mulaik (2009) describes additional indexes for special contexts, but I think many reviewers of submissions to journals would expect to see the minimal set:

(1) The *Steiger–Lind root mean square error of approximation* (RMSEA) and its 90% confidence interval (CI); in contrast to chi_M, which measures departure from perfect fit, the RMSEA measures departure from approximate fit in a correlation metric that also controls for N and df_M. *Approximate fit* means that chi_M does not exceed its expected value, df_M, over random samples when the model is true in the population. The best result is RMSEA = 0.

(2) The *Bentler comparative fit index* (CFI) compares the relative departures from approximate fit of the researcher's model compared with that of a baseline model in a standardized metric in which CFI = 1.0 is the best result.

(3) The *standardized root mean square residual* (SRMR) approximately measures the average absolute discrepancy between sample correlations and those predicted by the researcher's model for every pair of measured variables. The best result is SRMR = 0.

Values of global fit statistics computed in lavaan for the example analysis are:

$$\begin{array}{l} \text{chi}_\text{M}(7) = 7.320,\ p = 0.396 \\ \text{RMSEA} = 0.017, 90\%\ \text{CI}[0, 0.098] \\ \text{CFI} = 0.999, \text{SRMR} = 0.042 \end{array}$$

The model passes the chi-square test at the 0.05 level. However, the sample size is small, and the power of the chi-square test, in this analysis for N = 166 estimated in semTools, is only 0.18. This means that, if the model does *not* have perfect fit in the population, there is only a likelihood of 0.18 that this status will be detected in the chi-square test. Although RMSEA = 0.017 is not a terrible result, the upper bound of its 90% CI, or 0.098, is very close to 0.10, or so high that it signals *possible* ill fit at the level of the residuals. The result CFI = 0.999 is not alarming, and it says that the model in Figure 25.1 reduces the relative amount of departure from approximate fit by nearly 100% compared with a null model that assumes the endogenous variables are independent of each other and the exogenous variables. The result for the SRMR says that the average absolute difference between sample correlations and those predicted by the model is about 0.042; this is not a terrible result, but it masks problems at the level of the residuals.

Local Fit

Yamaga et al. (2013) did not describe the residuals in their original analysis, but we consider these results computed in lavaan for the same model and data. Reported in the top part of Table 25.1 are *correlation residuals* – differences between observed and predicted correlations for every pair of observed variables. They are continuous measures of local model–data discrepancies, and their values are relatively unaffected by sample size. Absolute correlation residuals > 0.10 signal a potential problem (Kline, 2023; Tabachnick & Fidell, 2013). It is hard to say exactly how many absolute correlation residuals ≥ 0.10 is too many, but the more there are, the worse the local fit. In Table 25.1, two absolute correlations shown in boldface exceed 0.10. For example, the correlation residual for the variables retention and oral health problems is −0.116. The sample correlation is −0.309 (Yamaga et al., 2013, p. 714), so the model underpredicts their association by −0.116 (i.e., the predicted correlation is −0.193). The path model in Figure 25.1 has no direct effect between these two variables, so perhaps that specification is an error (among other possibilities). The absolute correlation residual for ridge form and oral health problems, −0.114, is also relatively high (see Table 25.1).

The bottom part of Table 25.11 shows the *standardized residuals* – significance tests in the form of normal deviates (z) of the corresponding *covariance (raw, unstandardized) residuals*, or differences between sample and predicted covariances. Because covariances reflect the raw score metrics of both variables, it can be difficult to interpret the meaning of covariance residuals. Standardized residuals are more straightforward in their interpretation: If $z > 1.96$ in absolute value, then the corresponding covariance residual differs significantly from zero at the 0.05 level. In small samples, the power of standardized residuals is probably low. Nevertheless, the standardized residual for the pair retention and oral health problems (−2.214) is significant at the 0.05 level, and the result for the pair ridge form and oral health problems (−1.814) is nearly so (Table 25.1). Overall, there are signs of problematic fit at the level of the residuals, and any enthusiasm about global model fit should be

Table 25.1 *Correlation residuals and standardized residuals for a path model of denture satisfaction and oral health*

Variable	1	2	3	4	5	6	7
			Correlation residuals				
1. Jaw relation	0						
2. Mastication	0	0					
3. Satisfaction	0	0.009	0.007				
4. Oral health	−0.038	−0.004	−0.007	0.004			
5. Stability	0	0	0.051	−0.071	0		
6. Retention	0	0	0.071	**−0.116**	0	0	
7. Ridge form	0	0.080	0.033	**−0.114**	0	0	0
			Standardized residuals				
1. Jaw relation	0						
2. Mastication	0	0					
3. Satisfaction	0	1.195	1.195				
4. Oral health	−0.805	−1.195	−1.195	1.195			
5. Stability	0	0	0.892	−1.314	0		
6. Retention	0	0	1.228	**−2.124**	0	0	
7. Ridge form	0	1.195	1.195	−1.814	0	0	0

tempered here by knowledge of relatively poor explanatory power for certain pairs of variables in the model at the level of the residuals.

Equivalent Models

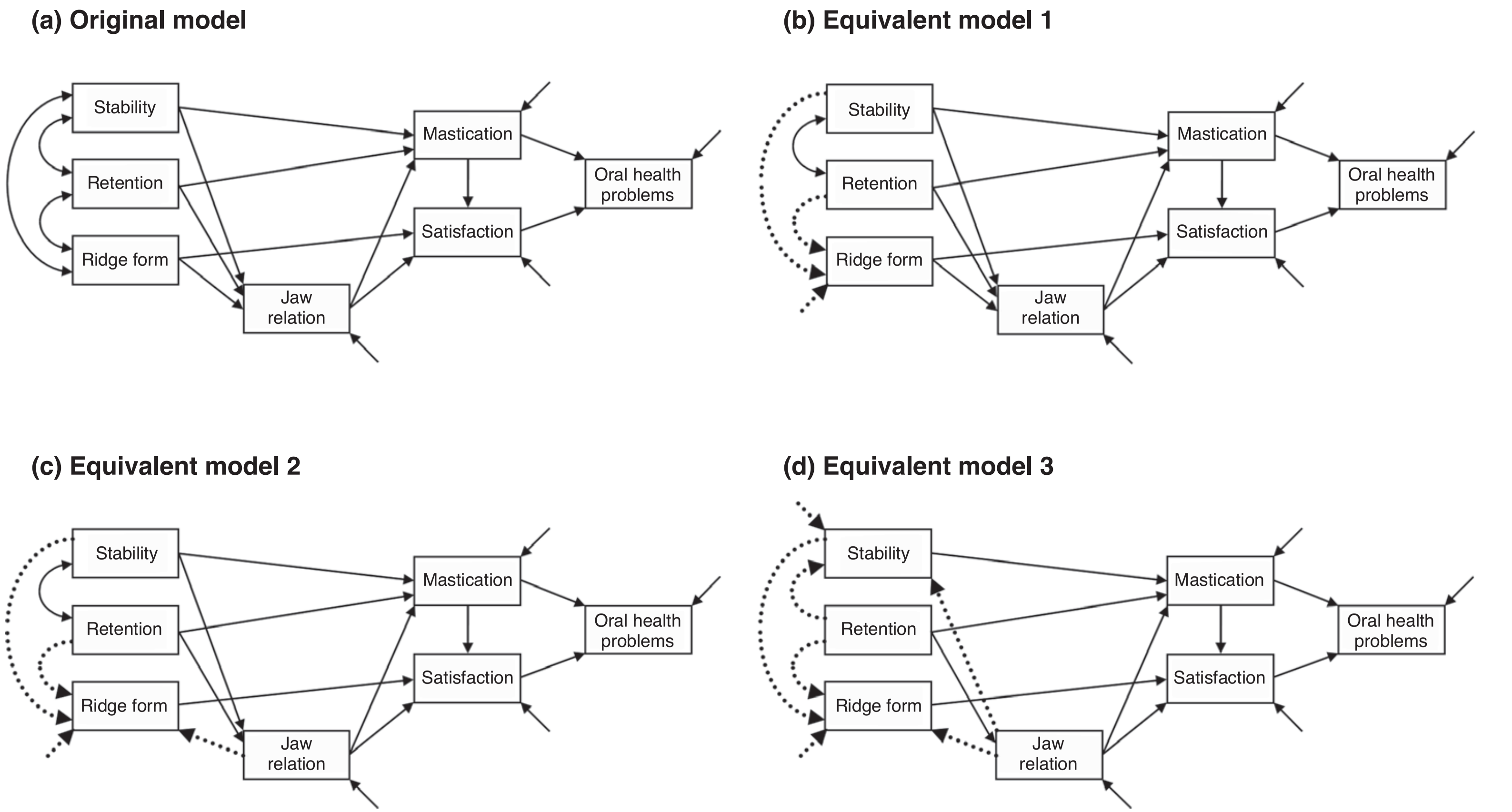
Next, we consider equivalent models. Figure 25.4(a) shows is the original Yamaga et al. (2013) path model for which chi_M (7) = 7.320. The other three models in the figure are equivalent versions generated by the *replacing rules*; they permit the substitution or reversal of certain paths *without affecting model fit* (Williams, 2012). For Figures 25.4 (a)–4(d), chi_M (7) = 7.320, and values of all other fit global fit statistics and residuals are exactly equal, but the equivalent models in Figures 25.4(b)–4(d) make opposing causal claims. For example,

(1) the status of the ridge form variable in Figure 25.4(b) is changed from exogenous, or causal in the original model, to endogenous, or an outcome in this equivalent version
(2) the direct causal effect between the ridge form and jaw relation variables is revered in Figure 25.4(c) compared with the original model
(3) the stability variable is specified as endogenous in Figure 25.4(d) – stability was exogenous in the original model
(4) the direct effect between stability and jaw relation is reversed in Figure 25.4(d) compared with Figure 25.4(a).

More equivalent versions of the original path model could be generated, so the variations in Figure 25.4 are not exhaustive. Yamaga et al. (2013) did not address the issue of equivalent models, a common shortcoming in SEM studies. A best practice would be for researchers to acknowledge the existence of at least a few plausible equivalent models and then argue why the original version is preferred. For example, Kale et al. (2000) retained a model of conflict resolution and relational capital, generated an equivalent version with identical fit, and gave arguments for their preferred model over the equivalent version. This level of transparency in SEM is commendable but rare.

Summary

The SEM family of techniques is flexible, used in many different areas, and can test a wide range of hypotheses about observed or latent variables. However, there are downsides to its increasing use in the social and behavioral sciences. This is especially true regarding incomplete reporting of the results, such as neglecting to describe full details about model fit. Respecting formal reporting standards for SEM would help to reduce incomplete reporting. Another common shortcoming is the failure to acknowledge the existence of equivalent models that explain the data just as well as the researcher's model. There are many additional kinds of models and analyses that are possible in SEM, but all require good judgment in their application and open, transparent reporting.


(a) Original model

(b) Equivalent model 1

(c) Equivalent model 2

(d) Equivalent model 3

Figure 25.4 *Original Yamaga et al. (2013) path model (a) and equivalent versions (b–d) all with identical fit to the data; dotted lines changed causal status relative to original model.*

Appendix

Syntax in lavaan for specifying and analyzing the example path model

```
# yamaga et al. (2013) path model
date()
options(”width” = 130)
library(lavaan)
library(semTools)
citation(”lavaan”, auto = TRUE)
citation(”semTools”, auto = TRUE)
# input data (covariances)
yamagaLower.cov <-'
  1.2769000
  0.1957612 0.5041000
  0.2490746 0.5287512 1.1449000
  0.3366496 0.2143064 0.3166772 0.9604000
  6.9383130 6.1569780 7.8776610 9.0972420 846.810000
  9.1743570 5.8470630 8.6872230 12.4053300 439.494390 846.810000
-4.4097120–2.9956320–4.7610720 -4.9815360–266.928480 -265.252320 207.360000 '
# add variable names
yamaga.cov <- getCov(yamagaLower.cov, names = c(“ridgeform”,"stability”,
  “retention”,"jawrelation”,"mastication”,“satisfaction”,
  ”oralhealth”))
# display covariances
yamaga.cov
# specify path model
  yamaga.model <-'
  jawrelation ~ stability + retention + ridgeform
  mastication ~ stability + retention + jawrelation
  satisfaction ~ ridgeform + jawrelation + mastication
  oralhealth ~ satisfaction + mastication '
# fit model to data, N = 166
yamaga.lavaan <- sem (yamaga.model, sample.cov = yamaga.cov,
  sample.nobs = 166)
summary(yamaga.lavaan, fit.measures = TRUE, standardized = TRUE,
  rsquare = TRUE)
# predicted covariance matrix
fitted(yamaga.lavaan)
# unstandardized, standardized, and correlation residuals
residuals(yamaga.lavaan, type = “raw”)
residuals(yamaga.lavaan, type = “standardized”)
residuals(yamaga.lavaan, type = “cor.bentler”)
# power of the chi-square test
findRMSEApower(0, .05, 7, 166, .05, 1)
```

References

Appelbaum, M., Cooper, H., Kline, R. B., et al. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board Task Force report. *American Psychologist*, *73*(1), 3–25. https://doi.org/10.1037/amp0000191

Arbuckle, J. L. (2021). *IBM SPSS Amos 28 User's Guide*. Amos Development Corporation.

Astrachan, C. B., Patel, V. K., & Wanzenried, G. (2014). A comparative study of CB-SEM and PLS-SEM for theory development in family firm research. *Journal of Family Business Strategy*, *5*(1), 116–128. https://doi.org/10.1016/j.jfbs.2013.12.002

Bagozzi, R. P. & Yi, Y. (2012). Specification, evaluation, and interpretation of structural equation models. *Journal of the Academy of Marketing Science*, *40*(1) 8–34. https://doi.org/10.1007/s11747-011-0278-x

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, *42*(5), 815–824. https://doi.org/10.1016/j.paid.2006.09.018

Bentler, P. M. & Wu, E. J. C. (2020). EQS 6.4 for Windows [Computer software]. Available at: https://mvsoft.com/.

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research*, 2nd ed. Guilford Press.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd ed. Erlbaum.

Deng, L., Yang, M., & Marcoulides, K. M. (2018). Structural equation modeling with many variables: A systematic review of issues and developments. *Frontiers in Psychology*, *9*, Article 580. https://doi.org/10.3389/fpsyg.2018.00580

Fan, Y., Chen, J., Shirkey, G., et al. (2016). Applications of structural equation modeling (SEM) in ecological studies: An updated review. *Ecological Processes*, *5*(1), Article 19. https://doi.org/10.1186/s13717-016-0063-3

Filippetti, V. A. & Krumm, G. (2020). A hierarchical model of cognitive flexibility in children: Extending the relationship between flexibility, creativity and academic achievement. *Child Neuropsychology*, *26*(6), 770–800. https://doi.org/10.1080/09297049.2019.1711034

Hayduk, L. A. (2014). Shame for disrespecting evidence: The personal consequences of insufficient respect for structural equation model testing. *Medical Research Methodology*, *14*(1), Article 124. https://doi.org/10.1186/1471-2288-14-124

Henley, A. B., Shook, C. L., & Peterson, M. (2006). The presence of equivalent models in strategic management research using structural equation modeling: Assessing and addressing the problem. *Organizational Research Methods*, *9*(4), 516–535. https://doi.org/10.1177/1094428106290195

Hoyle, R. H. & Isherwood, J. C. (2013). Reporting results from structural equation modeling analyses in *Archives of Scientific Psychology*. *Archives of Scientific Psychology*, 1, 14–22. https://doi.org/10.1037/arc0000004

JASP Team (2022). JASP (Version 0.16.1) [Computer software]. Available at: https://jasp-stats.org/

Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & J. S. Lang (eds.), *Testing Structural Equation Models* (pp. 294–316). SAGE Publications.

Jöreskog, K. G. & Sörbom, D. (1976). *LISREL III: Estimation of Linear Structural Equation Systems by Maximum Likelihood Methods*. National Educational Resources.

Jöreskog, K. G. & Sörbom, D. (2021). LISREL 11 for Windows [Computer software]. Available at: https://ssicentral.com/.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2022). semTools: Useful tools for structural equation modeling (R package 0.5-6). Available at: https://CRAN.R-project.org/package=semTools.

Kale, P., Singh, H., & Perlmutter, H. (2000). Learning and protection of proprietary assets in strategic alliances: Building relational capital. *Strategic Management Journal*, *21*(3), 217–237. https://doi.org/10.1002/(SICI)1097-0266(200003)21:3<217::AID-SMJ95>3.0.CO;2-Y

Kenny, D. A. & Milan, S. (2012). Identification: A nontechnical discussion of a technical issue. In R. H. Hoyle (ed.), *Handbook of structural equation modeling* (pp. 145–163). Guilford Press.

Kline, R. B. (2023). *Principles and Practice of Structural Equation Modeling*, 5th ed. Guilford Press.

Kühnel, S. (2001). The didactical power of structural equation modeling. In R. Cudeck, S. du Toit, & D. Sörbom (eds.), *Structural Equation Modeling: Present and Future. A Festschrift in Honor of Karl Jöreskog* (pp. 79–96). Scientific Software International.

Lang, K. M. & Little, T. D. (2018). Principled missing data treatments. *Prevention Science*, *19*(3), 284–294. https://doi.org/10.1007/s11121-016-0644-5

Little, T. D. (2013). *Longitudinal Structural Equation Modeling*. Guilford Press.

McDonald, R. P. & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*(1), 64–82. https://doi.org/10.1037/1082-989X.7.1.64

Mulaik, S. A. (2009). *Linear Causal Modeling with Structural Equations*. CRC Press.

Müthen, L. K. & Müthen, B. O. (1998–2017). *Mplus User's Guide*, 8th ed. Muthén & Muthén.

Boker, S., Nerale, M., Maes, H., et al. (2023). OpenMx: The OpenMx statistical modeling package. (R package 2.20.7). Available at. https://CRAN.R-project.org/package=OpenMx.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellora, D. T. (2018). The preregistration revolution. *PNAS*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Pek, J. & Hoyle, R. H. (2016). On the (in)validity of tests of simple mediation: Threats and solutions. *Social and Personality Psychology Compass*, *10*(3), 150–163. https://doi.org/10.1111/spc3.12237

Raykov, T. & Marcoulides, G. A. (2006). *A First Course in Structural Equation Modeling*, 2nd ed. Erlbaum.

Recio, L. A., Martín, P., Carvajal, F., Ruiz, M., & Serrano, J. M. (2013). A holistic analysis of relationships between executive function and memory in Parkinson's disease. *Journal of Clinical and Experimental Neuropsychology*, *35*(2), 147–159. http://dx.doi.org/10.1080/13803395.2012.758240

Rigdon, E. E. (2012). Rethinking partial least squares path modeling: In praise of simple methods. *Long Range Planning*, *45*(5–6), 341–358. https://doi.org/10.1016/j.lrp.2012.09.010

Rosseel, Y., Jorgensen, T. D., & Rockwood, N. (2022). lavaan: Latent variable analysis (R package 0.6-11). Available at: https://CRAN.R-project.org/package=lavaan.

SAS Institute Inc. (2021). *SAS/STAT 15.2 User's Guide*. SAS Institute Inc.

Sauvé, G., Kline, R. B., Shah, J. L., et al. (2019). Cognitive capacity similarly predicts insight into symptoms in first- and multiple-episode psychosis. *Schizophrenia Research*, *206*, 236–243. https://doi.org/10.1016/j.schres.2018.11.013

Shah, R. & Goldstein, S. M. (2006). Use of structural equation modeling in operations management research: Looking back and forward. *Journal of Operations Management*, ***24***(2), 148–169. https://doi.org/10.1016/j.jom.2005.05.001

StataCorp LLC (1985–2021). *Stata Structural Equation Modeling: Release 17*. Stata Press.

Steiger, J. H. (2001). Driving fast in reverse: The relationship between software development, theory, and education in structural equation modeling. *Journal of the American Statistical Association*, *96*(453), 331–338. https://doi.org/10.1198/016214501750332893

Tabachnick, B. G. & Fidell, L. S. (2013). *Using Multivariate Statistics*, 6th ed. Pearson.

Tarka, P. (2018). An overview of structural equation modeling: Its beginnings, historical development, usefulness and controversies in the social sciences. *Quality & Quantity*, *51*(1), 313–354. https://doi.org/10.1007/s11135-017-0469-8

Teo, T. (2010). A case for using structural equation modelling (SEM) in educational technology research. *British Journal of Educational Technology*, *41*(5), 89–91. https://doi.org/10.1111/j.1467-8535.2009.00999.x

Textor, J., van der Zander, B., & Ankan, A. (2020). dagitty: Graphical analysis of structural causal models (R package 0.3-0.). Available at: https://CRAN.R-project.org/package=dagitty.

Thelwall, M. & Wilson, P. (2016). Does research with statistics have more impact? The citation rank advantage of structural equation modeling. *Journal of the Association for Information Science and Technology*, *67*, 1233–1244. https://doi.org/10.1002/asi.23474

Tomarken, A. J. & Waller, N. G. (2003). Potential problems with "well-fitting" models. *Journal of Abnormal Psychology*, *112*(4), 578–598. https://doi.org/10.1037/0021-843X.112.4.578

Westfall, P. H., Henning, K. S. S., & Howell, R. D. (2012). The effect of error correlation on interfactor correlation in psychometric measurement. *Structural Equation Modeling*, *19*(1), 99–117. http://dx.doi.org/10.1080/10705511.2012.634726

Williams, L. J. (2012). Equivalent models: Concepts, problems, alternatives. In R. H. Hoyle (ed.), *Handbook of Structural Equation Modeling* (pp. 247–260). Guilford Press.

Williams, T. C., Bach, C. C., Matthiesen, N. B., Henriksen, T. B., & Gagliardi, L. (2018). Directed acyclic graphs: A tool for causal studies in paediatrics. *Pediatric Research*, *84*(4), 487–493. https://doi.org/10.1038/s41390-018-0071-3

Wold, H. (1982). Soft modeling: The basic design and some extensions. In K. G. Jöreskog & H. Wold, (eds.), *Systems Under Indirect Observations: Part II* (pp. 1–54). North-Holland.

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety.

Educational and Psychological Measurement, *73*(6), 913–934. https://doi.org/10.1177/0013164413495237

Wright, S. (1920). The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs. *Proceedings of the National Academy of Sciences*, *6*(6), 320–332. https://doi.org/10.1073/pnas.6.6.320

von Oertzen, T., Brandmaier, A. M., & Tsang, S. (2015). Structural equation modeling with Ωnyx. *Structural Equation Modeling*, *22*(1), 148–161. https://doi.org/10.1080/10705511.2014.935842

Xia, Y. & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods*, 51(1),409–428. https://doi.org/10.3758/s13428-018-1055-2

Yamaga, E., Sato, Y., & Minakuchi, S. (2013). A structural equation model relating oral condition, denture quality, chewing ability, satisfaction, and oral health-related quality of life in complete denture wearers. *Journal of Dentistry*, *41*(8), 710–717. https://doi.org/10.1016/j.jdent.2013.05.015

Zhang, M. F., Dawson, J., & Kline, R. B. (2021). Evaluating the use of covariance-based structural equation modelling with reflective measurement in organisational and management research: A review and recommendations for best practice. *British Journal of Management*, *32*(2), 257–272. https://doi.org/10.1111/1467-8551.12415

26 Multilevel Modeling

D. Betsy McCoach, Anthony J. Gambino, and Sarah D. Newton

Abstract

This chapter provides a brief introduction to multilevel models, specifically organizational models, and should be accessible to researchers who are familiar with ordinary least-squares (OLS) regression (i.e., multiple regression models). OLS regression assumes independence of observations; however, the responses of people clustered within organizational units (e.g., schools, classrooms, hospitals, companies) are likely to exhibit some degree of relatedness. In such scenarios, violating the assumption of independence produces incorrect standard errors that are smaller than they should be – multilevel modeling can alleviate this concern. However, the advantages of multilevel modeling are not purely statistical. Substantively, researchers may seek to understand the degree to which people from the same cluster are similar to each other and identify variables that predict variability within and across clusters. Multilevel analyses allow us to exploit the information in clustered samples and partition variance in the outcome variable into between-cluster and within-cluster variability. We can also use predictors at both the individual (level 1) and group (level 2) levels to explain this between- and within-cluster outcome variance.

Keywords: Multilevel Modeling, Hierarchical Linear Modeling, Random Coefficients Models, Organizational Models, Mixed Effects Models, Intraclass Correlation Coefficient

Introduction

Multilevel models are often referred to as *hierarchical linear*, *mixed*, *mixed-effects*, or *random-effects models*. Researchers use these terms interchangeably, although there are slight differences in their meanings. For instance, *hierarchical linear model* is a more circumscribed term that assumes a normally distributed outcome variable. In contrast, mixed-effects or random-effects models are more general than hierarchical linear models – they denote non-independence within a data set, but that non-independence does not necessarily need to be hierarchically nested. In this chapter, we focus on one type of random-effects model – the multilevel model – in which observations are hierarchically nested within higher-level structures. Specifically, we focus on *organizational models* – cross-sectional multilevel models where individuals (level-1 units) are clustered within an organizational, administrative, social, or political hierarchy (level-2 units).

Nested Data and Non-independence

Most traditional statistical analyses (e.g., ordinary least-squares [OLS] regression or multiple regression) assume that observations are independent (i.e., residuals are uncorrelated). However, the responses of people clustered within organizational units (e.g., schools, classrooms, hospitals, companies) are likely to exhibit some degree of relatedness, dependence, or interdependency, given that they were sampled from the same institution. For instance, students who attend the same school tend to be more similar in their achievement (and other educational outcomes) than students who attend different schools (Raudenbush & Bryk, 2002). In such scenarios, violating the assumption of independence produces incorrect standard errors that are smaller than they should be. Therefore, subsequent inferential statistical tests feature inflated Type I error rates – they produce statistically significant effects more often than they should. Multilevel modeling techniques allow researchers to explicitly model the relatedness of observations within clusters. The standard errors from multilevel analyses account for the clustered nature of the data, resulting in more-accurate Type I error rates (Raudenbush & Bryk, 2002).

In general, multilevel techniques allow researchers to model multiple levels of a hierarchy simultaneously, partition variance across the levels of analysis, and examine relationships and interactions among variables that occur across the hierarchy. Generally, the levels of interest within a multilevel model depend on the phenomena under investigation and the posed research questions (Gully & Phillips, 2019). For example, in a study of student achievement, students are nested within classrooms, so students are level-1 units and classrooms are level-2 units. However, classrooms (level-2 units) are nested within schools (level-3 units), and schools may be nested within school districts (level-4 units). Although such organizational models could include three or more levels, for the remainder of the chapter, we confine our discussion to two-level models.

Traditional correlation- and regression-based approaches estimate the relationship between two variables. However, standard single-level analyses like multiple regression (which ignore the clustered/hierarchical nature of the data) assume the relationship between the variables is constant across the entire sample. Multilevel modeling allows the relationships among key substantive variables to randomly vary across clusters. For example, the relationship between socio-economic status (SES) and achievement may vary by school. In some schools, student SES may be a strong (positive) predictor of subsequent academic achievement; in other schools, SES and academic achievement may be completely unrelated (Raudenbush & Bryk, 2002). Additionally, in multilevel modeling, researchers can study relationships among variables that occur at multiple levels of the data hierarchy (as well as potential interactions among variables at multiple levels) while allowing relationships among lower-level variables to randomly vary by cluster. Multilevel modeling allows us to ask and answer more-nuanced questions than are possible within traditional regression analyses (McCoach, 2019).

For instance, imagine we want to study the relationships between students' prior reading ability, school SES, and students' subsequent reading achievement. The data are clustered: students are nested within schools. Prior reading ability is an individual-level (level-1) variable, and school SES is a cluster-level (level-2) variable. We might hypothesize that school SES moderates the effect of students' reading ability on students' reading achievement. In other words, the relationship between students' prior reading ability and students' subsequent reading achievement varies as a function of school SES. For example, perhaps in high-SES schools, the relationship between prior reading ability and subsequent reading achievement is stronger (more positive) than it is in low-SES schools. In a standard linear regression model, we can include an interaction between school SES and student ability to examine whether school SES moderates the effect of prior reading ability on reading achievement. The multilevel model also allows the slope of students' prior reading ability on reading achievement to randomly vary across schools, even after controlling for all school- and student-level variables in the model. If the ability/achievement slope randomly varies across schools, even after including school SES in the model, school SES cannot fully explain the between-school variation in the ability/achievement relationship. Perhaps other, unmeasured variables could help to explain why the relationship between ability and achievement is stronger (more positive) in some schools than others. Even if it is not possible to explain why the relationship between ability and SES varies across schools, just knowing that the correlation between ability and achievement is stronger in some schools than others has important policy implications.

As the preceding paragraphs highlight, multilevel models are incredibly useful for studying organizational contexts like schools, companies, or families. However, many other types of data exhibit dependence. For instance, multiple observations collected on the same person represent another form of nested data. Growth-curve and other longitudinal analyses can be reframed as multilevel models, in which observations across time are nested within individuals. The multilevel-modeling framework partitions residual variance into within-person residual variance and between-person residual variance. In such a scenario, between-person residual variance represents across-person variability in any randomly varying level-1 parameters of interest, such as the intercept (which we commonly center to represent initial status in growth models) and the growth slope. Within-person residual variance represents the variance of time-specific residuals – generally referred to as measurement error.

For the remainder of this chapter, we focus exclusively on cross-sectional organizational models. First, we describe a set of two-level multilevel models, beginning with the simplest model – one with no predictors. Then, we introduce random intercept models, followed by models that include both randomly varying slopes and intercepts.

Multilevel Model with No Predictors

How does a multilevel model with no predictors differ from a multiple regression model with no predictors? In a clustered sample, people within a given cluster are more similar to each other than to individuals from other clusters.

Therefore, the residuals for observations (r_is) tend to be correlated within clusters, but independent across clusters. Given that the residuals for observations within clusters co-vary, some of the variance in the dependent variable can be explained by cluster membership. An additional error term, u_{0j}, captures the portion of the residual variance that is explained by membership in cluster *j*. The residual for the intercept for each cluster (u_{0j}) represents the deviation of a cluster's intercept from the overall intercept. The u_{0j} term allows us to model the dependence of observations from the same cluster because u_{0j} is the same for every person within cluster *j* (Raudenbush & Bryk, 2002).

For simplicity, let's first assume there are no predictors at level 2. The level-2 equation is then $\beta_{0j} = \gamma_{00} + u_{0j}$. In multilevel modeling, we refer to these β_{0j}s as *randomly varying* intercepts. The randomly varying intercept – on the right-hand side of the level-1 equation (acting as a predictor of *Y*) – is now on the left-hand side of the level-2 equation (acting as an outcome variable). The intercepts (β_{0j}s) are predicted by an overall intercept, γ_{00}, and a level-2 residual (error), u_{0j}, which captures the deviation of cluster *j*'s predicted intercept, β_{0j}, from the overall intercept, γ_{00}. Each of the *j* clusters has its own level-2 residual, u_{0j}, that allows each cluster to have its own intercept (β_{0j}). Rearranging the level-2 equation so that $u_{0j} = \beta_{0j} - \gamma_{00}$, the level-2 residual ($u_{0j}$) is the difference between β_{0j} – the expected cluster mean for the outcome variable – and γ_{00} – the overall expected value on the outcome variable. Thus, the set of multilevel equations for a completely unconditional model is

$$\begin{aligned} Y_{ij} &= \beta_{0j} + r_{ij} \qquad (26.1) \\ \beta_{0j} &= \gamma_{00} + u_{0j} \end{aligned}$$

The subscript for γ_{00} contains no *i* or *j* terms, meaning that γ_{00} is *fixed*; there is only one value of γ_{00} – the overall intercept. Because we have no predictors, γ_{00} is also the predicted mean (average) on the outcome variable, *Y.* Notice that β_{0j} occurs in both equations. Therefore, substituting $\gamma_{00} + u_{0j}$ for β_{0j} produces one combined (or mixed) equation:

$$Y_{ij} = \gamma_{00} + u_{0j} + r_{ij} \qquad (26.2)$$

But what does this mean? Person *i* in cluster *j*'s score on *Y* (Y_{ij}) equals the expected (predicted) mean (γ_{00}) plus their cluster's deviation from the overall mean (u_{0j}), plus their deviation from their own cluster's mean (r_{ij}).

For a more-concrete example, imagine that, on average, people report spending five hours on the internet per day ($\gamma_{00} = 5$). Laura lives in a house where the average number of hours spent on the internet per day is three ($\beta_{0j} = 3$), but Laura herself spends four hours on the internet per day ($Y_{ij} = 4$). Conceptually, $Y_{ij} = \gamma_{00} + u_{0j} + r_{ij}$ for Laura would be $4 = 5 + (-2) + 1$. In a non-multilevel framework (with a non-clustered, simple random sample), the prediction equation for Laura would simply be $Y_i = \beta_0 + e_i$, or $4 = 5 + (-1)$. The single-level regression equation contains only one error term – Laura's deviation from the overall average (or predicted) score. In contrast, the multilevel regression equation contains two residuals – the deviation of Laura's household from the overall mean (which in this case is −2) and Laura's deviation from her household mean (which in this case is +1). So, the overall mean (the overall intercept or predicted score) is the same in the multilevel and single-level frameworks above. What differs is our treatment of the residual(s) (McCoach & Cintron, 2022).

Random Effects and Variance Components

Without predictors, each person's score on the dependent variable is composed of three elements: the expected mean (γ_{00}), the deviation of the cluster mean from the overall mean (u_{0j}), and the deviation of the person's score from his/her cluster mean (r_{ij}). In this equation, γ_{00} is a *fixed effect*: γ_{00} is the same for everyone. The u_{0j} term is called a *random effect* for the intercept because u_{0j} randomly varies across the level-2 units (clusters). In multilevel modeling, *fixed effects* are parameters that are fixed to the same value across all clusters (or individuals), whereas *random effects* differ (vary) across clusters (West et al., 2015).

Residual Variances in Multilevel Models

Multilevel models and standard regression models do not differ in terms of their fixed effects. However, they differ in terms of the complexity of their residual variance/covariance structures. This more-complex residual variance/covariance structure is at the heart of multilevel modeling. Therefore, understanding the meaning and utility of the random effects that we include in multilevel models is essential.

To account for the dependence/clustering, we break the residual into two pieces: u_{0j} and r_{ij}; u_{0j} captures the deviation of the cluster mean (intercept) from the overall mean (intercept), and r_{ij} captures the deviation of the individual's score from the mean for that individual's cluster. We then compute variances for each of these residuals. The variance of r_{ij}, σ^2, represents the within-cluster residual variance in the outcome variable, and the variance of u_{0j}, τ_{00}, represents the between-cluster residual variance in the outcome.

We also make several important assumptions related to the residual variance terms: (1) the set of u values is normally distributed with a mean of 0 and a variance of τ_{00}; (2) the set of r values is normally distributed with a mean of 0 and a variance of σ^2; and (3) the within-cluster residuals (r_{ij}s) and between-cluster residuals (u_{0j}s) are uncorrelated (Raudenbush & Bryk, 2002). This last assumption allows us to cleanly partition the variance in the outcome variable into within- and between-cluster variance components. Therefore, in the simplest unconditional model with no predictors, the total variance in the outcome variable ($\mathrm{Var}(Y_{ij})$) equals the sum of the between-cluster variance (τ_{00}) and the within-cluster variance σ^2 (McCoach & Cintron, 2022).

Intraclass Correlation Coefficient

Let's delve a bit farther into our partitioning of the total variability in the outcome variable into within-cluster variance and between-cluster variance. The degree to which people within the same cluster differ from the cluster average is *within-cluster variability* – or the (pooled) variability across people within the same cluster. Conceptually, *between-cluster variability* represents the variability in the cluster means and is analogous to aggregating data to the cluster level, computing means for each cluster, and then estimating how much the cluster means vary.

The *intraclass correlation coefficient (ICC)* describes how similar individuals are within clusters and how much they vary across clusters; it quantifies the degree of dependence (relationship) among units from the same cluster (Hox, 2010; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). So, the ICC measures the proportion of between-cluster variance (the total variability in the outcome variable that is explained by cluster membership). The ICC also provides an estimate of the expected correlation between two randomly drawn individuals from the same cluster (Hox et al., 2017). Of course, the degree of dependence also varies by outcome variable, and some outcome variables may not exhibit any discernible dependence (even though the observations are clustered). Therefore, we must compute the ICC separately for each outcome variable of interest.

Because the ICC is the proportion of the total variability in the outcome variable that can be explained by cluster membership, the calculation of the ICC (often symbolized as ρ, "rho") involves partitioning the total variability in the outcome variable into within-cluster variance (σ^2) and between-cluster variance (τ_{00}). To compute the ICC, we simply divide the between-cluster variability (τ_{00}) by the total variability ($\tau_{00} + \sigma^2$), as the following formula shows:

$$\rho = \frac{\tau_{00}}{\tau_{00} + \sigma^2} \tag{26.3}$$

A large ICC indicates that there is a large degree of similarity within clusters (σ^2 is small) and/or a large degree of variability across clusters (τ_{00} is large). An ICC of 1 indicates that all observations within a cluster are perfect replicates of each other and all variability lies between clusters – the within-cluster variance is 0. In contrast, an ICC of 0 indicates that observations within a cluster are no more similar to each other than observations from different clusters – the between-cluster variance is 0. The assumption of independence implies an ICC of 0 (McCoach & Cintron, 2022).

To recap, cluster means vary (*between-cluster variance*). People in the same cluster also vary (*within-cluster variance*), although two people from a single cluster differ less than two randomly selected people. The sum of the within- and between-cluster variances represents total variance in the outcome variable. And the ICC indicates the proportion of total variability explained by group membership.

So, returning to our example of internet usage, mean internet usage varies by house (*between-house variance*). People in the same house also differ in how much they use the internet (*within-house variance*, though two people from the same house differ less than two randomly selected people). The sum of the within- and between-house variances represents the total variance in internet usage. And an ICC of 1.00 suggests people from the same house all report the exact same internet usage (i.e., all variation occurs across houses), whereas an ICC of 0.00 implies that people from the same house are just as likely to report similar internet usage as people from different houses (living in the same house has no influence on people's reported internet usage).

Intercepts as Outcomes Models: Predicting Randomly Varying Intercepts

Because intercepts vary across clusters, we can build a regression equation at level 2 to try to explain the variation in these randomly varying intercepts. For instance, in our internet usage example, we could include household-level covariates such as internet quality or the average age in the household as level-2 covariates to predict between-cluster variance in households' daily internet usage. Raudenbush and Bryk (2002) refer to these as *means as outcomes models* because the level-2 model predicts differences in the intercepts across clusters (level-2 units). The level-2 covariates may help to explain why some households spend more time using the internet than others. However, level-2 variables can never explain *within*-cluster variance (i.e., household-level variables cannot explain why certain members of the family use the internet more or less than other family members). To explain within-cluster (level-1) variance, we need to include within-cluster (level-1) covariates.

Adding Level-1 Predictors

Now, let's consider a model in which there is one individual-level predictor. Imagine that we want to predict daily hours spent on the internet using family member age. We regress hours using daily internet usage (Y_{ij}) on age (X_{ij}). Our level-1 model is:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij}) + r_{ij} \tag{26.4}$$

Remember, in standard linear regression, the intercept is the predicted value on Y when all predictors are held constant at 0. Similarly, we interpret the intercept (β_{0j}) as the predicted mean internet hours in cluster j when X_{ij}(age) is 0. Because age is 0 at birth, the intercept is the expected amount of internet usage for a newborn infant. The slope β_{1j} (the effect of age on internet usage) can vary by cluster, just like β_{0j} does. If we allow β_{1j} to randomly vary by cluster, β_{1j} becomes an outcome variable in a level-2 equation and has its own residual term, u_{1j}. Equation (26.5) contains the multilevel model with a randomly varying intercept and a randomly varying slope.

$$Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij}) + r_{ij} \tag{26.5}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

In Equation (26.5), γ_{00} represents expected (predicted) number of hours on the internet when age = 0; γ_{10} represents the average effect of age on internet usage across the entire sample. If age is measured in years, we expect a γ_{10}-hour change in daily internet usage for every one unit change in age. The error term, u_{1j}, represents the difference between the average slope and cluster *j*'s slope. In our example, u_{1j} is the difference between house *j*'s age/internet usage slope and the overall age/internet usage slope. If the "effect" of age on internet usage does not vary across clusters, then all clusters should have the same (or very similar) age/internet usage slopes. In such a scenario, the value of u_{1j} for each cluster would be 0 (or near 0), and the variance of u_{1j} would also be approximately 0. If the slope is the same across all clusters (i.e., the slope does not vary across clusters), it is not necessary to estimate a randomly varying slope. Instead, we could estimate a model in which the intercept for internet usage randomly varies across clusters, but the age/internet usage slope remains constant across clusters. In that case, our model equations would be:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij}) + r_{ij} \tag{26.6}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

Again, using substitution to combine these level-specific equations produces the combined model shown in Equation (26.7). If the age/internet usage slope does not randomly vary across clusters, the combined model is simple. Substituting $\gamma_{00} + u_{0j}$ for β_{0j} and γ_{10} for β_{1j}, the mixed-format equation is:

$$Y_{ij} = \gamma_{00} + \gamma_{10}(X_{ij}) + u_{0j} + r_{ij} \tag{26.7}$$

Such that person *ij*'s score on *Y* is a function of γ_{00} (the overall intercept; the predicted score when $X_{ij} = 0$ when age = 0), γ_{10} – the slope of age on internet usage – multiplied by X_{ij} (person *ij*'s age), u_{0j} – the deviation of his/her household's predicted daily number of hours spent on the internet at age 0 from the overall intercept, and r_{ij} – the deviation of person *ij*'s score from his/her model predicted score.

As an aside, it is common to refer to "effects" in multilevel modeling (e.g., the effect of age on daily internet usage). In fact, the entire lexicon of the technique is replete with references to fixed effects, random effects, cross-level interaction effects, and so forth. However, these "effects" are not necessarily indicative of a causal mechanism. The causal claims that can be made from a multilevel analysis are determined by the strength of the research design, and multilevel analyses do not strengthen inferences obtained from weak designs (Kelloway, 1995).

Randomly Varying Slopes and Intercepts

If the age/internet usage slope does randomly vary by cluster, then substituting $\gamma_{10} + u_{1j}$ for β_{1j} results in the following combined equation:

$$Y_{ij} = \gamma_{00} + \gamma_{10}\left(X_{ij}\right) + u_{0j} + u_{1j}\left(X_{ij}\right) + r_{ij} \tag{26.8}$$

Person *ij*'s score is again a function of γ_{00}, γ_{10} multiplied by X_{ij}, u_{0j}, and r_{ij}, but it is also a function of u_{1j} (the deviation of his/her household's slope from the overall slope) multiplied by X_{ij} (person *ij*'s age).

Allowing the age/internet usage slope to randomly vary across households by including a *random effect* for the slope (u_{1j}) specifies a model in which the age/internet usage slope is different for different households. Therefore, in some households, there could be no relationship between age and internet usage, resulting in an age/internet usage slope of 0; in other households, the age/internet usage slope could be negative, indicating that older members of the household tend to use the internet less than younger members of the household. Finally, the age/ internet usage slope could be positive, indicating that older members of the household tend to use the internet more than younger members of the household. The fixed effect, γ_{10}, indicates the expected (average) value of the age/internet usage slope across the entire sample. The variance in the age/internet usage slope, Var(u_{1j}), indicates how much households vary from that overall average. A great deal of variance in u_{1j} indicates a lot of between-household variability in the age/internet usage slope. In contrast, if the variance of u_{1j} is 0, then there is no variability across households in terms of their age/internet usage slopes; in this case, we would want to fix u_{1j} to 0 to greatly simplify the model (McCoach & Cintron, 2022).

Full Contextual (Slopes-as-Outcomes) Model

The *full contextual model* contains both level-1 and level-2 predictors. Level-2 predictors may help to explain between-cluster differences in the intercept, the expected value of the outcome variable when all the level-1 variables are held constant at 0. Level-2 predictors may also help explain between-cluster differences in level-1 slopes. In other words, the level-2 variable helps to predict why the relationship between the level-1 predictor and the outcome variable differs across clusters. Returning to our example, age is a level-1 predictor of daily internet usage. We could include a household-level variable, such as the quality of the internet in the home, to predict the average number of hours spent on the internet within the household (the intercept). When a level-2 variable (i.e., the house's internet quality) is a predictor of a level-1 slope (i.e., the age/internet usage slope), we refer to this term as a *cross-level interaction* because it represents an interaction between a level-2 variable and a level-1 variable (McCoach & Cintron, 2022). In this example, perhaps the mean age of the household moderates the relationship between age and internet usage. For example, perhaps in younger households, with parents and small children, the relationship between age and internet usage is positive (i.e., older family members use the internet more than younger members). In contrast, in older households, such as intergenerational households with teenagers, middle-aged parents and grandparents, the relationship between age and internet usage is likely to be negative (i.e., the younger family members use the internet more than the older members.)

Cross-level interactions accomplish an important task – explaining random slope variance. Multilevel models without cross-level interactions still capture that a randomly varying slope differs across clusters. However, including cross-level interactions informs us about which variables predict/explain that slope variation. Equation (26.9) represents the multilevel model that includes a cross-level interaction between X_{ij} (individual *ij*'s age) and W_j (house *j*'s mean age).

$$Y_{ij} = \beta_{0j} + \beta_{1j}\left(X_{ij}\right) + r_{ij} \tag{26.9}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\left(W_j\right) + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}\left(W_j\right) + u_{1j}$$

Substituting the equations for β_{0j} and β_{1j} for those terms in the first equation results in the following combined equation:

$$Y_{ij} = \gamma_{00} + \gamma_{10}\left(X_{ij}\right) + \gamma_{01}\left(W_j\right) + \gamma_{11}\left(W_j\right)\left(X_{ij}\right) + u_{0j} + u_{1j}\left(X_{ij}\right) + r_{ij} \tag{26.10}$$

The cross-level interaction, $\gamma_{11}\left(W_j\right)\left(X_{ij}\right)$, appears exactly as an interaction would in a single-level model.

Variance/Covariance Components

The γ terms are the *fixed effects*, and the u terms are the *random effects*. All γ terms could be estimated using single-level regression models. However, the u terms, the random effects, are unique to mixed/multilevel models. Multilevel techniques allow us to model, estimate, and test the variances (and covariances) of these random effects (also known as *variance components* – denoted by the symbol τ_{qq}). Specifically, τ_{00} represents the variance of the randomly varying intercepts (u_{0j}), τ_{11} signifies the variance of the first randomly varying slope (u_{1j}), etc. In addition, we generally allow the random effects (within a given level) to co-vary with each other. Therefore, in our simple example above, τ_{01} represents the covariance between residuals for the randomly varying intercepts and slopes.

$$\text{Var}\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{01} & \tau_{11} \end{bmatrix} \tag{26.11}$$

Standardized τ_{01} represents the correlation between the residuals of the intercept and slope. If τ_{01} is positive, clusters with more positive intercepts also tend to have more positive (less negative) slopes. If τ_{01} is negative, clusters with more positive intercepts tend to have less positive (more negative) slopes. In our example, if τ_{01} is positive, it means that after controlling for the other variables in the model (i.e., household age and household SES), households that have higher internet usage also tend to have more positive age/internet usage slopes. Remember, in multilevel modeling, although random effects can co-vary within a given level, residuals are uncorrelated across levels. As a result, though u_{0j} and u_{1j} are allowed to co-vary, both u_{0j} and u_{1j} are uncorrelated with r_{ij} (Raudenbush & Bryk, 2002).

Advice on Modeling Randomly Varying Slopes

Depending on the researcher's theoretical framework and the sample size at level 1, the slopes for some of the level-1 predictors may randomly vary across level-2 units, or they may be fixed across all level-2 units. A random-coefficients model contains one or more randomly varying level-1 slopes (Raudenbush & Bryk, 2002). Although our simple example contains only one level-1 variable (age), multilevel models often contain several level-1 variables. We could easily include several level-1 control variables, such as gender, race/ethnicity (often a set of four to six dummy coded variables), free lunch status, English learner status, and special education status. In such a situation, the researcher must decide which level-1 slopes to allow to randomly vary across schools and which level-1 slopes to fix to a single value across all schools.

But why not allow all level-1 covariates to randomly vary across schools? First, remember the structure of the residual covariance matrix. The unstructured tau matrix (τ) contains a variance for the randomly varying intercept and every randomly varying slope, as well as all possible covariances among the slopes and intercept. Therefore, the number of unique variance/covariance components in the tau matrix is equal to $r(r+1)/2$, where r equals the number of random effects. With a random intercept and one random slope, the tau matrix contains $(2*3)/2=3$ parameters (two variances and a covariance). However, in a model that contains five randomly varying slopes and a randomly varying intercept, the tau matrix contains $(6*7)/2=21$ unique parameters; the tau matrix for a model with 10 randomly varying slopes and a randomly varying intercept contains $(11*12)/2=66$ unique parameters. In other words, in a model with 10 randomly varying slopes (and an unstructured tau matrix), we need to estimate a total of 67 different residual parameters: σ^2 and an 11 × 11 tau matrix containing 66 unique level-2 variance/covariance components. Partitioning the residual variance in a model into 67 separate pieces feels like a Sisyphean task (especially given that standard regression models estimate just one residual variance parameter).

Raudenbush and Bryk (2002) cautioned against succumbing to the "natural temptation to estimate a 'saturated' level-1 model" in which all level-1 predictors are specified to have randomly varying slopes (p. 256). Parsimony is a key consideration for several reasons (McCoach & Cintron, 2022):

(1) First, as demonstrated above, adding random slopes radically increases the complexity of the model.

(2) The number of random slopes is limited by the level-1 sample size. The number of level-1 units must exceed the number of variance components. An extreme example occurs in dyadic analysis. Because there are only two units within each cluster, it is only possible to estimate one random effect per cluster.

(3) It is common to experience convergence problems when trying to estimate randomly varying slopes that are unnecessary. Multilevel models that contain random slopes with no between-cluster variance often fail to converge (or require thousands of iterations to reach a solution). Because variances cannot be less than 0, trying to estimate randomly varying slopes that are actually 0 in the population often leads to boundary issues (and convergence issues). Unfortunately, such results may not provide guidance about which random effects to eliminate. (McCoach et al., 2018).

Therefore, we recommend being judicious and parsimonious about which random slopes to estimate in your multilevel models. Include randomly varying slopes if they are central to your research question or if you have compelling evidence from prior research that the slopes are likely to randomly vary. Eliminate any unnecessary random effects for level-1 coefficients that do not vary across level-2 units (McCoach et al., 2018). For example, we tend not to allow slopes for demographic covariates (e.g., race, gender, ethnicity, SES, age) to randomly vary when they serve as control variables in our models. However, if our research focused on understanding how the relationship between SES and achievement varies across schools and what school factors influence this relationship, we would certainly allow the SES slope to randomly vary across schools.

Centering Level-1 Predictors

In regression models, we often *center* covariates for both substantive and analytic reasons. As mentioned earlier, the intercept is the predicted value of the outcome variable when all predictor variables are held constant at 0. In our internet usage example, the intercept for internet usage was the predicted number of hours spent using the internet per day at age 0. In single-level regression, one common strategy is to *center* continuous predictor variables by subtracting the mean of the variable ($\overline{X}$) from each person's score (X_i). This transforms person *i*'s score on *X* into a deviation score indicating how far above or below the mean person *i* was. Therefore, the mean of a centered variable is 0 and the variance is the same as the variance of the score in its original metric (because all scores change only by a single constant value – the mean). In single-level regression, centering continuous covariates is especially important when including interaction terms. The choice of centering influences the main effects for the predictor variables included in the interaction term; the regression coefficient is the predicted effect of *X* on *Y* when the other predictor variable in the interaction term equals 0 (Aiken & West, 1991).

In multilevel modeling, we center continuous predictor variables for substantive and/or analytic reasons. First, centering continuous covariates allows for a more substantively useful and interpretable intercept. Second, the magnitude of the between-person (residual) variance in the intercept, τ_{00}, and the correlation between the intercept and any randomly varying slopes is dependent on the location of the intercept. In organizational applications of multilevel modeling, the two main centering techniques for lower-level covariates are *grand-mean centering* and *group-mean centering*. Grand-mean centering subtracts the overall mean of the variable from all scores. Therefore, the grand-mean-centered score captures a person's standing relative to the full sample. Group-mean centering subtracts the cluster's mean from each score in the cluster. As such, the transformed score captures a person's standing relative to their own cluster.

In our example, let's imagine grand-mean and group-mean centering age (X_{ij}) for person i in cluster j. The grand mean represents the mean age across all individuals i and all households j ($\overline{X}..$) in the entire sample, whereas the cluster mean represents the mean age across all individuals i in a given household j ($\overline{X}._j$). To grand-mean center age, we subtract the mean age in the entire sample from each person ij's age ($X_{ij} - \overline{X}..$), so X_{ij} is person ij's deviation from the overall average age ($\overline{X}..$). To group-mean center age, we subtract the average age in household j from the age of each member of household j ($X_{ij} - \overline{X}._j$), so X_{ij} is person ij's deviation from his/her household's average age ($\overline{X}._j$).

Obviously, the decision about how to center independent variables has major implications for the interpretation of the intercept. Grand-mean centering sets the intercept at the overall mean. This holds age constant at the overall mean, thereby controlling for age. When grand-mean centering age, the randomly varying intercept, β_{0j}, denotes the predicted number of hours using the internet for household j, assuming that this household's average age is the same as the overall average age in the sample. The overall intercept, γ_{00}, is the predicted number of hours using the internet, holding age constant at the overall mean ($\overline{X}..$). In this case, grand-mean-centered age represents each person's deviation from the average age across the entire sample. Grand-mean centering represents a simple linear transformation of the original variable.

One problem with grand-mean centering arises when no one in cluster j has scores near the overall mean. In such cases, the intercept for that cluster is extrapolated outside the range of data for the cluster. For example, if the average age across households is 40 but, in household j, the four members are 55, 55, 65, and 65 years old, the grand-mean-centered scores are 15, 15, 25, and 25. No one in the household has a centered score near 0. Thus, the intercept in household j is the predicted daily internet usage in hours for a 40-year-old, even though there are no 40-year-olds in that household. For a detailed discussion of the statistical and interpretational issues that such extrapolation can cause, see Raudenbush and Bryk (2002).

On the other hand, if we group-mean center age, the randomly varying intercept (β_{0j}) is the mean number of hours spent online in household j. Having subtracted each cluster's own mean ($\overline{X}._j$) from each score, the mean of the cluster-mean-centered age variable is 0 in every cluster; therefore, the randomly varying intercept (β_{0j}) for each cluster is the mean (expected/predicted) daily number of hours spent using the internet in that household (j). Group-mean-centered age represents each person's deviation from his/her own household's average age. So, in a household where the ages are 55, 55, 65, and 65, the household's mean age is 60. To group-mean center, we subtract 60 from each score, producing group-mean-centered scores of −5, −5, 5, and 5. The mean of the group-mean-centered variable is 0 in every cluster, so the overall intercept (γ_{00}) is the mean of cluster means: it is the overall average household daily internet usage.

Important Guidance on Group-Mean Centering

Group-mean centering removes between-cluster variation from the level-1 covariate, so the variance of a group-mean-centered variable provides an estimate of the pooled within-cluster variance (Enders & Tofighi, 2007). With group-mean centering, we can partition variance in the predictor, the outcome, and the relationship between the predictor and the outcome into within- and between-cluster components. However, group-mean centering does not preserve information about between-cluster differences on the X variable. Although the group-mean-centered score provides information on an individual's relative standing as compared to their respective cluster, it provides no information about the individual or the group's relative standing as compared to the overall sample. Using a different cluster mean to center each cluster results in a centered X variable that contains information about how much a person deviates from his/her group, but contains no information about how much the person deviates from the overall mean on X. For example, grand-mean-centered scores of 15, 15, 25, and 25 indicate that two members of the household are 15 years older than the sample average and two are 25 years older than the sample average. In contrast, the group-mean-centered scores of −5, −5, 5, and 5 tell us nothing about the how the ages in this household compare to ages in the other households in the sample.

Therefore, when group-mean centering, it is important to include the aggregate of the group-mean-centered variable (or a higher-level variable that measures the same construct) into the analysis. Without an aggregate or contextual variable at level 2, all the information about between-cluster variability in the X variable is lost. In our age example, the grand-mean-centered age's mean for our cluster, +20, provides information indicating that the average age in this cluster is 20 years older than the average age in the overall sample. In contrast, the cluster mean for our cluster (and every other cluster in the sample) is 0. However, adding the cluster mean into the model as a level-2 predictor preserves the between-cluster component of the age variable. Finally, when group-mean centering, a different cluster mean is subtracted from each cluster. Therefore, group-mean centering is not a simple linear transformation of the uncentered model, and it does not produce results that are statistically equivalent to the uncentered and/or grand-mean-centered results.

There is some debate within the multilevel literature about whether to use grand-mean centering or group-mean centering (Enders & Tofighi, 2007). Because centering decisions affect the interpretations of important model parameters involving the intercept, it is important to carefully and thoughtfully decide if and how to center covariates. The decision to use grand-mean or group-mean centering may vary, depending on the context of the study, the research questions asked, and the nature of the variables in question. For instance, if the primary research question involves understanding the impact of a level-2 variable on the dependent variable, and the level-1 variables serve as control variables, grand-mean centering may be an appropriate choice. On the other hand, when level-1 variables are of primary research interest, or for research on contextual and compositional effects, group-mean centering may be more appropriate (Enders & Tofighi, 2007). In addition, group-mean centering aids in the computation of variance explained (*R*-squared) measures (Rights & Sterba, 2019b). To preserve between-cluster information from the covariate, we recommend including the aggregates of any group-mean-centered variables at level 2.

What about centering level-2 variables? Grand-mean centering is the only available option at level 2. Generally, it is advisable to grand-mean center all level-2 continuous variables. When using level-2 variables as part of a cross-level interaction, grand-mean centering is especially important. However, even for level-2 variables that predict only randomly varying intercepts (not randomly varying slopes), grand-mean centering the level-2 variable usually facilitates interpretation of the intercept. When reporting multilevel modeling results, it is important to explain centering decisions and procedures and to interpret the parameter estimates accordingly. See Enders and Tofighi (2007) for an excellent discussion of centering in organizational multilevel models.

Model Adequacy and Fit

Now that we have spent some time thinking about constructing multilevel models, interpreting their parameters, and making methodological decisions (e.g., how to center predictors, whether to estimate random effects, etc.), what do we do with our resulting model? Our next steps focus on evaluating multilevel models in terms of model fit (i.e., how well did our model fit our data?) and model adequacy (i.e., how much variation in the outcome of interest did our model explain?).

Model Selection

We must always remember that statistical models are just representations of our data; none portray exact truth (Burnham & Anderson, 2004). As Box famously said, “All models are wrong, but some are useful” (Box & Draper, 1987, p. 424). So, how do we determine which model is best? Generally, three considerations guide the model selection process (Burnham & Anderson, 2004): (1) parsimony – estimating more parameters cannot worsen model fit (Forster, 2000), but consider whether improved model fit warrants adding extra parameters; (2) Comparison of multiple theoretical hypotheses is more informative than comparing a theoretical model to the often-implausible null hypothesis (Burnham & Anderson, 2004); and (3) evidence for specific theoretical models (Burnham & Anderson, 2004) may be more desirable than evidence against the atheoretical null model.

Criteria for Evaluating Model Fit

When assessing model/data fit in multilevel modeling, we often evaluate *nested* models – where one model is a subset of the second – with the likelihood ratio test (LRT or deviance difference test; Raudenbush et al., 2000). The deviance formula is: −2*log-likelihood. The null model (M_0) is more parsimonious, with deviance $= D_0$ and p_0 estimated parameters; the alternative model (M_1) is more parameterized, with deviance $= D_1$ and p_1 estimated parameters. The LRT compares the deviance difference ($\Delta D = D_0 - D_1$) to the critical value of χ^2, with $df = \Delta p = p_1 - p_0$. If ΔD exceeds the χ^2 critical value, this indicates statistically significantly improved model fit favoring the more-complex model (M_1). However, if M_1 fails to reduce the deviance substantially, we retain the more-parsimonious model, M_0. Therefore, statistically significant deviance decreases support the more-complex model, whereas non-significant decreases support the less-complex model (McCoach et al., 2022). However, traditional LRTs are less appropriate for testing boundary parameters, such as random effects (Berkhof & Snijders, 2001), which cannot be normally distributed around a mean of 0 if the null hypothesis is true (Dominicus et al., 2006; Stoel et al., 2006). To test such boundary parameters, the correct critical value comes from the $\bar{\chi}^2$ distribution, instead of the typical χ^2 distribution (Snijders & Bosker, 2012).

Additionally, information criteria (ICs), such as the Akaike Information Criterion (AIC; Akaike, 1973) and Bayesian Information Criterion (BIC; Schwarz, 1978), assess *nested* or *non-nested* models, evaluate goodness of fit, and quantify the strength of data-based evidence for each model (Burnham et al., 2011). ICs feature two components: the deviance (*D*; i.e., model fit) and some per-parameter penalty on the deviance that accounts for model complexity – the number of estimated parameters (*p*) and/or the sample size (*n*; Dominicus et al., 2006). We favor models with lower ICs.

$$\mathrm{AIC} = D + 2p \tag{26.12}$$

$$\mathrm{BIC} = D + \ln(n) * p \tag{26.13}$$

Unfortunately, there is no definitive consensus as to which sample size (e.g., total sample size, number of clusters, effective sample size) to use to calculate the BIC (Skrondal & Rabe-Hesketh, 2004). Regardless, due to the magnitudes of their per-parameter penalties, AIC tends to favor more-parameterized models; BIC tends to favor less-parameterized models. Thus, the AIC, BIC, and LRT may prefer different models. Therefore, in addition to model fit, it is important to examine measures of model adequacy, such as those presented in Rights and Sterba's (2019b) multilevel modeling variance decomposition framework.

Variance Decomposition Framework for Multilevel Modeling

Rights and Sterba (2019b, p. 309) developed "an integrative framework of R-squared measures for MLMs [multilevel models] with random intercepts and/or slopes based on a completely full decomposition of variance." This framework decomposes the model-implied total variance into five sources of variation, variance attributable to: level-1 predictors via fixed slopes (f_1); level-2 predictors via fixed slopes (f_2); level-1 predictors via random slope variation and covariation (v); cluster-specific outcome means via random intercept variation (m); and level-1 residuals (σ^2). All level-1 predictors must be group-mean-centered to achieve this full decomposition. Otherwise, the variance attributable to predictors via fixed slopes is represented by a single source (f) for all levels. For simplicity, this chapter assumes all level-1 predictors were group-mean centered. Decomposing the model-implied total variance into these five sources enables the computation of variance-explained measures and provides potential insights into the model's predictive capability – its degree of model adequacy.

The f_1, v, and σ^2 sources contain only *within*-cluster (co)variation, and their sum is the model-implied within-cluster variance. Therefore, we can evaluate the proportion of within-cluster variance explained by each (or a combination) of these sources. Similarly, the f_2 and m sources contain only *between*-cluster (co)variation, and their sum is the model-implied between-cluster variance. Therefore, we can evaluate the proportion of between-cluster variance from these sources individually or in combination. Finally, we can evaluate the proportion of total variance explained by each (or a combination) of these five sources, which sum to the model-implied total variance.

Understanding Proportion of Variance-Explained Measures

Calculating the R^2 measures described in Rights and Sterba (2019b) requires the model parameter estimates and the sample variance/covariance matrix for the predictors included in the multilevel model. Given this information, the following formulae produce Rights & Sterba's (2019b) proposed R^2-measure components:

$$f_1 = \boldsymbol{\gamma}_w' \boldsymbol{\Phi}_w \boldsymbol{\gamma}_w \tag{26.14}$$

$$f_2 = \boldsymbol{\gamma}_b' \boldsymbol{\Phi}_b \boldsymbol{\gamma}_b \tag{26.15}$$

$$v = tr(\mathbf{T\Sigma}) \tag{26.16}$$

$$m = \tau_{00} \tag{26.17}$$

where $\boldsymbol{\gamma}_w$ is the vector of level-1 fixed-effect slope estimates, $\boldsymbol{\Phi}_w$ is the variance/covariance matrix of the level-1 predictors, $\boldsymbol{\gamma}_b$ is the vector of level-2 fixed-effect slope estimates, $\boldsymbol{\Phi}_b$ is the variance/covariance matrix of the level-2 predictors, $tr()$ is the trace function, $\mathbf{T}$ is the variance/covariance matrix of the random effects, $\mathbf{\Sigma}$ is the variance/covariance matrix of the level-1 predictors with randomly varying slopes (this includes a variance of 0 for the intercept and covariances of 0 between the

Table 26.1 *Description of* R^2 *measures based on the Rights and Sterba (2019b) framework*

R^2 Measure	Interpretation
$R_w^{2(f_1)} = \frac{f_1}{f_1 + v + \sigma^2}$	The proportion of model-implied *within-cluster* variance explained by the *fixed effects of level-1 predictors*.
$R_w^{2(v)} = \frac{v}{f_1 + v + \sigma^2}$	The proportion of model-implied *within-cluster* variance explained by the *(co)variation of the random effects of level-1 predictors*.
$R_b^{2(f_2)} = \frac{f_2}{f_2 + m}$	The proportion of model-implied *between-cluster* variance explained by the *fixed effects of level-2 predictors*.
$R_b^{2(m)} = \frac{m}{f_2 + m}$	The proportion of model-implied *between-cluster* variance explained by the *variation of the random intercept*.
$R_t^{2(f_1)} = \frac{f_1}{f_1 + f_2 + v + m + \sigma^2}$	The proportion of model-implied *total* variance explained by the *fixed effects of level-1 predictors*.
$R_t^{2(f_2)} = \frac{f_2}{f_1 + f_2 + v + m + \sigma^2}$	The proportion of model-implied *total* variance explained by the *fixed effects of level-2 predictors*.
$R_t^{2(v)} = \frac{v}{f_1 + f_2 + v + m + \sigma^2}$	The proportion of model-implied *total* variance explained by the *(co) variation of the random effects of level-1 predictors*.
$R_t^{2(m)} = \frac{m}{m + f_2 + f_1 + v + \sigma^2}$	The proportion of model-implied *total* variance explained by the *variation of the random intercept*.

intercept and all of the predictors because the intercept is a constant), and τ_{00} is the random intercept variance estimate. Finally, the σ^2 component is the level-1 residual variance estimate.

Using these estimated sources, we can compute variance-explained measures to assess model adequacy (see Table 26.1). Additionally, any measures that refer to the same type of variance (i.e., within-cluster, between-cluster, or total, represented by the subscripts *w*, *b*, and *t*, respectively) can be summed to produce a measure of the proportion of that type of variance explained by the corresponding sources together.

For example, imagine that we estimated the model shown in Equation (26.18) for our daily internet usage outcome, where X_{ij} is each individual's age, X_j is each house's average age (the cluster means for age), and W_j is each house's internet quality. Table 26.2 displays a fabricated set of values for this example and their interpretations.

$$\begin{aligned} Y_{ij} &= \beta_{0j} + \beta_{1j}\left(X_{ij} - \overline{X}_{\cdot j}\right) + r_{ij} \qquad (26.18) \\ \beta_{0j} &= \gamma_{00} + \gamma_{01}\left(W_j - \overline{W}_{\cdot}\right) + \gamma_{02}\left(X_j - \overline{X}_{\cdot}\right) + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11}\left(W_j - \overline{W}_{\cdot}\right) + u_{1j} \end{aligned}$$

Table 26.2 *Example* R^2 *measures for internet usage model*

R^2 Measure	Interpretation
$R_w^{2(f_1)} = 0.500$	(The fixed effect of) age explained 50% of internet usage's within-house variance.
$R_w^{2(v)} = 0.100$	10% of the within-house variance in internet usage is explained by age's random effect variation.
$R_b^{2(f_2)} = 0.200$	(The fixed effects of) household internet quality and average age explained 20% of internet usage's between-house variance.
$R_b^{2(m)} = 0.800$	Internet usage's random intercept variation explained 80% of its between-house variance.
$R_t^{2(f_1)} = 0.375$	(The fixed effect of) age explained 37.5% of internet usage's total variance.
$R_t^{2(f_2)} = 0.050$	(The fixed effects of) household internet quality and average age explained 5% of internet usage's total variance.
$R_t^{2(v)} = 0.075$	Age's random effect variation explained 7.5% of internet usage's total variance.
$R_t^{2(m)} = 0.200$	Internet usage's random intercept variation explained 20% of its total variance. This is equivalent to a conditional ICC, so 20% of the unexplained variance is between-house variance.

Effect Size

An *effect size* is a practical, interpretable, quantitative measure of the magnitude of an effect. As with any statistical analyses, it is important to report effect size measures for multilevel models. The R^2 measures described above can help researchers and readers to determine the impact that a variable or a set of variables has on a model, with respect to variance explained. In addition, researchers can compute Cohen's *d*-type effect sizes to describe the mean differences among groups. To calculate the equivalent of Cohen's *d* for a group-randomized study (where the treatment variable occurs at level-2), use the following formula (Spybrook et al., 2006):

$$\delta = \frac{\hat{\gamma}_{01}}{\sqrt{\hat{\sigma}^2 + \hat{\tau}_{00}}} \tag{26.19}$$

Assuming two groups have been coded as 0/1 or −0.5/+0.5, the numerator of the formula utilizes $\hat{\gamma}_{01}$ (the model estimate of γ_{01}), which represents the difference between the treatment and control groups. The denominator utilizes $\hat{\sigma}^2$ and $\hat{\tau}_{00}$ (the model estimates of σ^2 and τ_{00}) from the unconditional model, where the total variance in the dependent variable is divided into two components: the between-cluster variance, τ_{00}, and the within-cluster variance, σ^2. There are numerous ways to compute effect sizes in multilevel modeling (or any analysis), and not all effect sizes need to be standardized, especially when unstandardized metrics are commonly used and/or easily understood. Ultimately, the goal is to present the results of themodel as clearly as possible and to contextualize the parameters in practically meaningful and easily interpretable ways.

Model Building Steps for Confirmatory/Predictive Models

Next, we provide a recommended sequence for building and testing confirmatory multilevel models, using Rights and Sterba's (2019b) integrative proportion of variance explained framework and the model fit criteria outlined above.

Step 1: First, fit the unconditional random-effects analysis of variance (ANOVA) model (*Model 1*). This is the model with no predictors. The main goal of fitting this model is to compute and report the unconditional ICC for the outcome variable. This unconditional model has only three parameters: the within-cluster variance, the between-cluster variance, and the parameter estimate for the intercept – the overall expected value on the outcome variable.

Step 2: Fit the full theoretical model, including all within- and between-cluster variables, all hypothesized cross-level (and same-level) interactions, and all theoretically relevant random effects (*Model 2*). Using Rights and Sterba's (2019b) framework to partition the variances into the five components mentioned above requires *group-mean centering of all* level-1 variables; generally, it is best practice to include the aggregates of those group-mean-centered variables at level 2. Carefully consider which level-1 slopes are allowed to randomly vary at level 2. If theory specifies a cross-level interaction between a level-2 variable and a level-1 variable, add that interaction to the model, even if the level-1 slope is not randomly varying. *Therefore, the decision to allow for a random slope should be based on your hypothesis about whether the slopes will randomly vary AFTER accounting for slope variability attributable to cluster-level variables* (LaHuis & Ferguson, 2009).

Step 3: (Optional/not always necessary). After running the full contextual model, if any random effects prove to be unnecessary, eliminate them. Re-estimate the model, and compare the model fit and proportions of variance explained within clusters, between clusters, and overall for the simpler model to the full contextual model (Model 2). However, be cautious – determining whether to eliminate random effects for one or more slopes is not completely straightforward. We recommend examining the following four pieces of information to gauge whether any of the randomly varying slopes are unnecessary:

(i) Model convergence issues are often a sign that the multilevel model contains an unnecessary random slope. If you experience model convergence problems, or if it takes thousands of iterations for the model to converge, you may need to eliminate one or more random effects.

(ii) Some software packages provide tests of the statistical significance for the variance components. You can consult these tests for guidance; however, statistical tests of variance components are somewhat controversial and should be treated as approximations or heuristics to provide guidance, not exact and infallible tests. Also, in programs that usually report standard errors, their absence for some or all the variance components is often a sign that one or more of the random effects in the model is unnecessary.

(iii) Use Rights and Sterba's (2019b) framework to compute the proportion of within-cluster (outcome) variance explained by the level-1 predictors via random slope variation/covariation in the model that includes the randomly varying slope(s). If the model contains multiple randomly varying slopes, compare the $R_w^{2(v)}$ from the model that includes all of the slopes as randomly varying to the $R_w^{2(v)}$ from the model that constrains one of the slopes to be non-random. If the change in R^2 is near 0.00, allowing the slope to randomly vary explains very little of the within-cluster variance, suggesting that the random slope could potentially be eliminated.

(iv) Compare the model fit (deviance, AIC, BIC) of the model that includes the random slope to the model that does not. The AIC tends to be more liberal than the BIC or the χ^2 difference (LRT) test (with a small number of degrees of freedom) and suggests a penalty of 2.00 points per parameter. So, deviance changes of less than two points per eliminated parameter suggest that model fit is very similar across the two models; the fit of the more-parsimonious model and the fit of the more-complex model (the one that includes the random slope) are similar. In such cases, we favor the simpler model. Therefore, deviance differences of less than two points per parameter suggest eliminating the random slope, but remember that eliminating a random slope does not necessarily reduce the model by only one parameter. Fitting an unstructured variance/covariance matrix for the level-2 variance components allows for covariances among all the level-2 variance components. Therefore, a model with one random slope and one random intercept has three variance/covariance parameters, whereas a model with only a random intercept has one variance/covariance parameter. Thus, the model with the random slope contains two additional parameters, and, according to the AIC, the deviance should drop by at least four points to justify including the additional parameter (McCoach & Cintron, 2022).

Step 4: (Optional). Sometimes researchers compare a model that eliminates one or more fixed effects from the model to the full contextual model. The most common rationale for fitting the reduced model and comparing it to the full model is to compute a change in R-squared measure, providing a method for determining how much variance is uniquely explained by the variable eliminated from the reduced model. This change in R^2 provides an indication of the predictive utility of the predictor. Rights and Sterba (2019a) provide a detailed demonstration of their framework to compare their R^2 measures across models.

Step 5: Report the results of your analyses. When reporting the results of your analyses, be sure to explain the centering/coding for each of the predictor variables. Also, thoroughly describe both the fixed effects and random effects included in each of the models and describe the structure of the variance/

covariance components. The methodology and results should provide a detailed description of the entire taxonomy of estimated multilevel models, as well as justifications for any decisions to trim or modify your models. Results tables should include the coefficients for the fixed effects, the level-1 and level-2 variance components, the deviance, and the number of estimated parameters for each of the multilevel models. Be sure to describe the practical significance and include effect size measures to help the reader to interpret the practical magnitude of your results. For an in-depth example of this process, see McCoach et al. (2022).

Conclusion

Multilevel modeling is a powerful tool for researchers working with data structures that naturally feature dependence among observations. Multilevel modeling solves certain statistical issues that arise from non-independent or clustered data, and it allows for more-nuanced analyses of variables that occur at different levels of data hierarchies. Nevertheless, its application requires careful thought and attention, and it cannot necessarily solve design issues that threaten the validity of causal claims.

Although this chapter provides a solid conceptual overview of the technique, several important areas remain unaddressed, including residual analysis, power and sample size issues, three-level models, and modeling heterogeneity of the level-1 residual variances. To learn more about multilevel modeling, we recommend consulting the following books: Raudenbush and Bryk (2002), Hox et al., (2017), Goldstein (2011), O'Connell et al. (2022), and/or Snijders and Bosker (2012).

References

Aiken, L. S. & West, S. G. (1991). *Multiple Regression: Testing and Interpreting Interactions*. SAGE Publications.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & B. F. Csaki (eds.), *Second International Symposium on Information Theory* (pp. 267–281). Academiai Kiado.

Berkhof, J. & Snijders, T. A. B. (2001). Variance component testing in multilevel models. *Journal of Educational and Behavioral Statistics*, *26*(2), 133–152. https://doi.org/10.3102/10769986026002133

Box, G. E. P. & Draper, N. R. (1987), *Empirical Model-Building and Response Surfaces*. John Wiley & Sons.

Burnham, K. P. & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*(2), 261–304. https://doi.org/10.1177/0049124104268644

Burnham, K. P., Anderson, D. R., & Huyvaert, K. P. (2011). AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. *Behavioral Ecological Sociobiology*, *65*, 23–35. https://doi.org/10.1007/s00265-010-1029-6

Dominicus, A., Skrondal, A., Gjessing, H. K., Pedersen, N. L., & Palmgren, J. (2006). Likelihood ratio tests in behavioral genetics: Problems and solutions. *Behavior Genetics*, *36*(2), 331–340. https://doi.org/10.1007/s10519-005-9034-7

Enders, C. K. & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*, 121–138. http://dx.doi.org/10.1037/1082-989X.12.2.121

Forster, M. R. (2000). Key concepts in model selection: Performance and generalizability. *Journal of Mathematical Psychology*, *44*(1), 205–231. https://doi.org/10.1006/jmps.1999.1284

Goldstein, H. (2011). *Multilevel Statistical Models (Kendall's Library of Statistics 3)*, 4th ed. Edward Arnold.

Gully, S. M. & Phillips, J. M. (2019). On finding your level. In S. E. Humphrey & J. M. LeBreton (eds.), *The Handbook of Multilevel Theory, Measurement, and Analysis* (pp. 11–38). American Psychological Association. https://doi.org/10.1037/0000115-002

Hox, J. J. (2010). *Multilevel Analysis: Techniques and Applications*, 2nd ed. Routledge.

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2017). *Multilevel Analysis: Techniques and Applications*, 3rd ed. Routledge.

Kelloway, E. K. (1995). Structural equation modeling in perspective. *Journal of Organizational Behavior*, *16*, 215–224.

LaHuis, D. M. & Ferguson, M. W. (2009). The accuracy of significance tests for slope variance components in multilevel random coefficient models. *Organizational Research Methods*, *12*(3), 418–435. https://doi.org/10.1177%2F1094428107308984

McCoach, D. B. (2019). Multilevel modeling. In G.R. Hancock, L. M. Stapleton, & R. O. Mueller (eds.) *The Reviewers Guide to Quantitative Methods in the Social Sciences* (pp. 292-312), 2nd ed. Routledge.

McCoach, D. B. & Cintron, D. W. (2022). *An Introduction to Modern Modeling Methods*. SAGE Publications.

McCoach, D. B., Rifenbark, G. G., Newton, S. D., et al. (2018). Does the package matter? A comparison of five common multilevel modeling software packages. *Journal of Educational and Behavioral Statistics*, *43*(5), 594–627. https://doi.org/10.3102/1076998618776348

McCoach, D. B., Newton, S., & Gambino, A. J. (2022). Evaluating the fit and adequacy of multilevel models. In A. A. O'Connell, D. B. McCoach, & B. A. Bell (eds.), *Multilevel Modeling Methods with Introductory and Advanced Applications*. Information Age Publishing.

O'Connell, A. A., McCoach, D. B., & Bell, B. A. (eds.), *Multilevel Modeling Methods with Introductory and Advanced Applications.* Information Age Publishing.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd ed. SAGE Publications.

Raudenbush S., Bryk, A., Cheong, Y. & Congdon, R. (2000). *HLM Manual*. SSI International.

Rights, J. D. & Sterba, S. K. (2019a). New recommendations on the use of *R*-squared differences in multilevel model comparisons. *Multivariate Behavioral Research*, *55*(4), 568–599. https://doi.org/10.1080/00273171.2019.1660605

Rights, J. D. & Sterba, S. K. (2019b). Quantifying explained variance in multilevel models: An integrative framework for defining *R*-squared measures. *Psychological Methods*, *24*(3), 309–338. https://doi.org/10.1037/met000018

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://www.jstor.org/stable/2958889

Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman & Hall/CRC Press.

Snijders, T. A. B. & Bosker, R. J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, 2nd ed. SAGE Publications.

Spybrook, J., Raudenbush, S. W., Liu, X. F., Congdon, R., & Martínez, A. (2006). Optimal design for longitudinal and multilevel research: Documentation for the “Optimal Design” software. Survey Research Center of the Institute of Social Research at University of Michigan.

Stoel, R. D., Garre, F. G., Dolan, C., & van den Wittenboer, G. (2006). On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychological Methods*, *11*(4), 439–455. https://doi.org/10.1037/1082-989X.11.4.439

West, B. T., Welch, K. B., & Galecki, A. T. (2015). *Linear Mixed Models: A Practical Guide Using Statistical Software*, 2nd ed. Routledge.

27 Meta-Analysis

Yuri Jadotte, Anne Moyer, and Jessica Gurevitch

Abstract

Meta-analysis is a form of data synthesis that statistically combines the results of primary research studies responding to a given question. It has become an indispensable tool for decision making and advancement of knowledge in a variety of disciplines. This chapter provides an overview of this method, beginning with a brief discussion of systematic reviews – the research methodology that undergirds meta-analysis. The chapter then explores specific components of this approach as it is most widely applied in the literature, including issues related to effect sizes, heterogeneity of study outcomes, scope of the analysis, and quality-control issues to consider when conducting a meta-analysis. A brief overview of new and emerging methods for the synthesis of primary research data is also provided, highlighting different forms of meta-analysis and different approaches for the synthesis of research data. Practical examples are provided as illustrations to clarify and reinforce the concepts presented in this chapter.

Keywords: Evidence Synthesis; Systematic Review; Scientific Generality; Evidence-Based Decisions; Scoping Review; Research Synthesis; Effect Size; Vote Counting Procedure

Introduction

Meta-analysis is the statistical synthesis of the results of different studies addressing similar questions. It is used very broadly across many fields (e.g., medicine, psychology, education, criminal justice, epidemiology, nursing, clinical trials, ecology, evolution, and conservation). Meta-analysis is a relatively new statistical field, founded in the last quarter of the twentieth century. Meta-analysis is part of the broader field of research synthesis that includes systematic reviews. There are four core reasons for doing a meta-analysis: (1) to resolve discrepancies in the outcomes of a group of studies; (2) to highlight where more research is needed or where evidence is already sufficient; (3) to avoid biases, providing replicability and transparency in reviewing research evidence; and (4) to allow broader generalization by encompassing a broader scope than is possible in any one study.

Meta-analyses can range from small and narrowly focused to very large, broad in scope, and general. How large should yours be? The “Goldilocks” just-right-size meta-analysis depends on the questions you are asking of the literature, the size of the literature (i.e., number of studies), the scientific discipline in which you are working, and your resources for conducting the meta-analysis. One does not want to undertake a meta-analysis that is so large that it will take 25 years to complete – it would be outdated long before it was done, and one's spouse and friends would probably never speak to you again. One doesn't want to do a meta-analysis that is so small that it answers few questions, and the results apply in only very narrow circumstances. If you conduct an initial search and find that only three studies have been published on the topic, you might want to broaden the topic and increase the scope. If the initial search reveals hundreds of thousands of studies, you should refine or narrow the scope.

So how do you know if you've chosen the right number of studies and the right scope for your meta-analysis? First, decide whether you want to do a systematic review or scoping review. A systematic review (see Chapter 4 in this volume) is the research methodology that undergirds the collection, evaluation, and reporting of the data that go into a meta-analysis. Unlike traditional narrative reviews, systematic reviews are a scientific approach to collecting and evaluating the literature on a question and are transparent and aim to be repeatable. The two essential steps in a systematic review are a thorough and unbiased search to find all the relevant evidence on the question addressed, and an assessment of whether the evidence can be included in the meta-analysis, depending on its quality and other a priori considerations. These steps must be undertaken before statistically combining data using meta-analysis to draw conclusions about the effectiveness of a program, policy, or intervention. A systematic review requires an a priori literature search strategy, with specific, replicable search terms, and criteria for selection of studies recorded in advance and reported in a protocol following standard guidelines such as PRISMA (http://prisma-statement.org; O'Dea et al., 2021). Searches are conducted using multiple scientific literature databases (e.g., Scopus, Web of Science, Medline). Unpublished data (e.g., dissertations, government reports, other "gray literature") may be included. Study search and screening procedures, and the assessment of the risk of bias within each study, should ideally be carried out by two people to minimize human error. Scoping reviews use the same methods for finding and selecting studies (Peters et al., 2015), but they are generally more narrative and are used to map existing evidence and identify gaps, particularly when there is a very large number of primary studies (Snilstveit et al., 2016).

Once a systematic review is conducted, and the relevant studies are identified from the literature, a meta-analysis can be undertaken. We begin by discussing effect sizes – these put the results of different studies on the same scale. We continue with statistical models for combining studies, evaluating heterogeneity among the studies, and considering moderators (covariates) that may explain differences among studies. We then deep dive into important (if subtle) issues that can determine the quality of the meta-analysis and conclude with insights on new and emerging methods for synthesizing research data that expand upon or even transcend conventional meta-analytic statistical approaches.

Meta-Analysis Effect Sizes

Imagine a wedding for a couple whose guests hail from several different countries. Their cash gifts include US dollars, euros, Turkish lira, Canadian dollars, and Swiss francs. How might one tell who gave the most lavish gifts, so that one could write them an especially gracious thank you note? How might one understand, on average, how generous the guests have been? These questions could be answered by converting all cash gifts to a *common currency* to rank and properly average them. This is the logic of using effect sizes to summarize the findings of studies addressing a similar research question but using different measures. Effect sizes are indices that allow the magnitude and direction of study findings to be synthesized and modeled across studies.

A key step lies in correctly identifying the effect size metric to use to summarize the study outcomes. One commonly used effect size is the standardized mean difference, which characterizes continuous outcome measures in units of standard deviations (e.g., calories, mass, height, reading scores, level of symptoms, degree of functioning). Standardized mean differences are calculated as the difference between the means of two groups, such as an experimental or intervention group and a control group, divided by their pooled standard deviation. This effect size indicates by how many standard deviations the target group scores better (or worse) than the control or comparison group. A value of 0.0 indicates no difference between the groups. Expressing the difference in units of standard deviations is similar to a *t*-test and provides information about the effect when the assumptions of normally distributed data are met.

Other common effect sizes include a variety of ratios. A commonly used effect size in ecology and some other disciplines is the log response ratio – the natural log of the ratio of the means of two groups. For example, metabolic rate might be 20% higher in an experimental group compared to a control. In medical studies, and other fields where outcomes are often dichotomous (e.g., survived/died, suffered/did not suffer a myocardial infarction, remained out of prison/returned to prison, reproduced/did not reproduce), the odds ratio is often an appropriate effect size. “Odds” are the probability of something happening divided by the probability of it not happening – an odds ratio of 1.00 indicates no effect. They are usually analyzed as log odds ratios to normalize their distribution. The relative risk is a similar measure used when incidence data are available. A hazard ratio is used when incidence rate data are available to account for duration of exposure to a given risk or factor for each participant in a study. In studies that assess the association between two continuous measures (e.g., adolescents’ hours of screen time per day and hours of sleep per night), correlation coefficients are typically used. They are transformed using Fisher’s *Z*-transformation to make them normally distributed.

Poor data reporting, particularly in older literature, can make calculating effect sizes difficult. For example, whereas odds ratios and correlation coefficients are commonly reported in primary research studies, standardized mean differences and response ratios must often be calculated by the meta-analyst digitizing data in graphical form presented in research reports, using a program such as ImageJ (https://imagej.nih.gov/ij/), or calculated from information in online supplements. It is becoming more common in some fields for effect sizes to be published. Other metrics of effect size can also be reliably aggregated across studies, particularly if their statistical properties are known and sample sizes are large.

Sometimes, a study reports results that are opposite or "inverse" to the values of other studies (e.g., most studies examine vitality, but some report fatigue). In other cases, the experimental manipulation is the reverse or inverse of that in other studies (some experiments report a response to a reduction in crowding, while most of the studies report a response to increases in crowding). A convention is for scores on "opposite" measures of an outcome, or for opposite treatments, to be reverse-scored so that higher scores indicate comparable responses, allowing them to be meaningfully combined. Most important is to be clear about what you are measuring and to align that with the question(s) you are asking. When there is no obvious categorization for the two groups being compared (they do not fit neatly into an intervention and a control, or the "control" is not comparable among studies), the meta-analyst needs to be thoughtful in determining which one is the "baseline." Imagine we are concerned about responses to crowding in experimental studies, where studies compare responses in high-density versus lower-density environments. The meta-analysis must be performed such that "high density" is consistently being compared to "low density" across all studies, regardless of how they are characterized in the primary studies.

To account for the impact of moderators in a meta-analysis (discussed below), values for moderators are assigned for each study. For example, characteristics shared by participants (e.g., class size, participant ages, gender, income, study duration, intensity of intervention, or other meaningful variables) may be recorded for each study included in the meta-analysis as study-level variables.

Aggregating Effect Sizes Across Studies

The basic premise of meta-analysis is that summarized data for each outcome (each of the effect sizes – analogous to the dependent variable in primary experimental studies) are combined across multiple primary research studies. Larger studies are assumed to better estimate the "true effect" of a treatment or intervention because they have smaller sampling variances and thus higher precision than smaller studies. This is taken into account in combining results across studies and analyzing them appropriately. Perhaps you want to know whether a certain innovative educational approach increases reading scores. In one study, 75 individuals are taught using this novel approach and are compared with 75 others who are taught using the conventional approach. In another study, 5,000 individuals are exposed to the novel approach and compared to 5,000 individuals who have the conventional educational approach. Naturally, we would like to count the large-study results more heavily, because very small studies are more likely to be subject to chance outcomes.

In meta-analysis, we can calculate the sampling variance – the variance in estimating the effect size if we were to do the same study many times – of an effect size if we know the statistical properties of the effect size metric. We then use that to weigh larger, more precise estimates of the outcome more heavily than smaller, less precise estimates. Every study is assumed to estimate some “true” outcome. If the sample size used in the study was large enough, we would get close to that true effect of the intervention.

If we weigh studies only by their sampling variance, we are assuming that the only reason the studies’ outcomes differ from one another is the precision with which they estimate the true outcome. This conceptual approach is formalized statistically in what is called a fixed effect model. We may not believe that all studies share a common effect, though. We can add another component to our statistical model by incorporating an additional variance term to account for random differences in the true effects among studies. This is called a random effects model. In a random effects model, we assume that variations in the effect among the studies is due to both sampling variance within each study and true variance in the outcomes among studies. Fixed effect meta-analyses tend to have smaller confidence intervals and are more likely to falsely reject the null hypothesis. While early meta-analyses used fixed effect models, random effects models for meta-analyses are now generally preferred in most research disciplines.

The weighted effect sizes are then averaged across studies with the appropriate equations (Borenstein et al., 2009). Meta-analysis results are typically presented numerically using a mean effect size (e.g., the mean relative risk, odds ratio, or standardized mean difference across studies), with the appropriate 95% confidence interval. In addition, heterogeneity statistics and *p*-values indicate whether the mean effect differs from the value for “no effect.” The grand mean effect size across studies is “statistically significant” if its confidence interval does not overlap with the numerical value of “no effect” – that is zero for a standardized mean effect size or 1.0 for an odds ratio or relative risk. In addition, the results of a meta-analysis are almost always presented in a graphical format known as the *forest plot* (Figure 27.1), which visually represents the effect size magnitude and precision for individual studies, and the overall effect size and precision across all studies pooled together in the meta-analysis. The mean effect size may tell us a lot about the response to the treatment or intervention, or it may not be very meaningful, depending on the heterogeneity among studies (i.e., how different the studies are from each other). These concepts are examined further in the next sections.

Examining Heterogeneity: Variance Between Studies and the Role of Moderators

The variability in the true effect across a set of studies is called heterogeneity. Each study differs from the others in various ways, such as their methodology and the characteristics of their populations, specifics of the interventions, and comparators (e.g., control treatments). We can choose to lump those unknown differences – above and beyond sampling variation – into the heterogeneity term. If there is a great deal of heterogeneity among studies, we would interpret that to mean that the true effect – the real response to the intervention – differs among studies. If the heterogeneity is small, the outcomes are largely in agreement among

Studies	Estimate (95% CI)	Ev/Trt	Ev/Ctrl
ALLHAT 2002	0.905 (0.792, 1.034)	380/5170	421/5185
Asselbergs 2004	1.884 (0.807, 4.397)	15/431	8/433
GISSI 2008	0.873 (0.622, 1.224)	61/2285	70/2289
Kyushu Lipid Intervention Study Group 2000	1.227 (0.743, 2.028)	40/2219	24/1634
Nakamura 2006	0.528 (0.295, 0.947)	17/3866	33/3966
Pravastatin Multinational Study Group 1993	0.077 (0.004, 1.367)	0/530	6/532
Ridker 2008	0.456 (0.298, 0.696)	31/8901	68/8901
Sasaki 2002	0.326 (0.117, 0.909)	5/587	13/498
Sever 2003	0.624 (0.478, 0.815)	86/5168	137/5137
Shepherd, Cobbe 1995	0.700 (0.580, 0.844)	174/3302	248/3293
Shepherd, Blauw 2002	0.846 (0.732, 0.978)	299/2891	356/2913
Stegmayr 2005	0.695 (0.261, 1.852)	6/70	9/73
Wanner 2005	0.853 (0.663, 1.097)	93/619	112/636
Yusuf 2016	0.650 (0.448, 0.945)	45/6361	69/6344
Amarenco 2006	0.669 (0.524, 0.855)	101/2365	151/2366
Colhoun (CARDS trial) 2004	0.654 (0.463, 0.924)	51/1428	77/1410
Domanski 2007	0.372 (0.171, 0.809)	6/74	207/950
Downs 1998	0.599 (0.433, 0.830)	57/3304	95/3301
Fellstrom 2009	0.959 (0.857, 1.074)	415/1389	431/1384
Heart Protection Study Group 2005	0.738 (0.595, 0.916)	141/10269	191/10267
Heljic 2009	0.476 (0.131, 1.732)	3/45	7/50
Holdaas (ALERT trial) 2003	0.754 (0.570, 0.997)	79/1050	105/1052
Knopp (ASPEN trial) 2006	0.728 (0.500, 1.058)	45/959	61/946
Koizumi (Holicos-PAT trial) 2002	0.893 (0.597, 1.336)	58/1422	37/810
Overall (I^2 = 57.22 %, P< 0.001)	**0.749 (0.679, 0.825)**	**2208/64705**	**2936/64370**

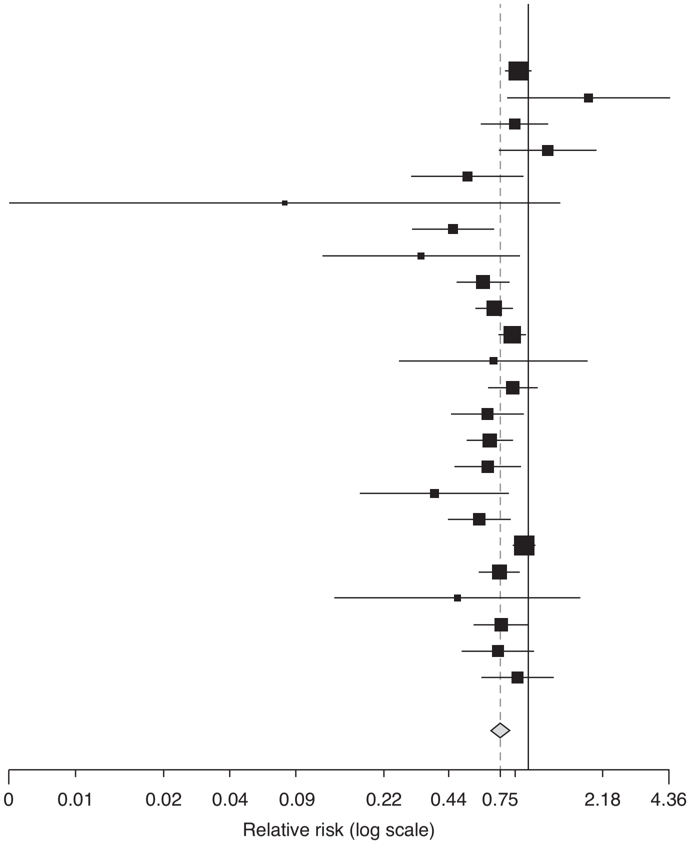


Figure 27.1 *Example of a forest plot and tabular display of data. Symbol sizes indicate weights. CI, confidence interval; Ctrl = control group; Ev = number of events; Trt = treatment group.*

the studies. To the extent that results are heterogeneous, the mean effect size across studies may be less likely to reliably characterize the responses across studies. It may not be useful for providing sound recommendations for practice or policy.

Indices of heterogeneity help specify whether the variation in the mean effect across studies is reliably different from zero. A *Q*-test (Hedges & Olkin, 1985) is a statistical assessment used to determine whether such excess variation is present by chance alone, and the I^2 statistic is used to estimate the proportion of the total variation in true effect sizes that is due to heterogeneity in the real responses across studies (Higgins & Thompson, 2002). For example, as shown in Figure 27.1, the I^2 value signifies that 57.22% of the total variance among the studies in the meta-analysis is due to true differences between the studies and not just to sampling variance or chance alone. Generally, I^2 values of 25, 50, and 75 are thought of as low, moderate, and high heterogeneity, respectively, although there is no consensus on the exact thresholds to use. Why does it matter if there is high heterogeneity in the responses among studies? Some of this heterogeneity may not be due to random variation but to factors that can be identified – moderators (called covariates or explanatory variables in different research fields). These categorical or continuous explanatory factors might explain some of the variation or heterogeneity among a set of effect sizes. Moderators might include, for example, the age or educational level of the participants, the duration of each study, or the environment in which the studies were conducted. Moderators are coded at the study level and help answer the questions of for whom, under what conditions, and using what procedures and measures, the effects are larger versus smaller.

If the questions being asked in the meta-analysis are about the main effects (e.g., whether this social intervention has the desired outcome), the moderators can help to reduce artefacts (e.g., how the experiments were conducted). In some research areas, the main effects are not of primary interest or are not especially meaningful because researchers may be more interested in hypothesized factors that modify that effect; for example, what characteristics of the intervention program or of the populations studied determine the effectiveness of interventions to prevent recidivism in perpetrators of domestic violence. In that case, the moderators (e.g., different kinds and durations of interventions, different characteristics of the perpetrators) are of particular interest in the meta-analysis. This is especially the case in ecological meta-analyses. Reviews whose main purpose is to draw firm conclusions about a specific intervention or treatment may implement more restrictive inclusion/exclusion criteria, to reduce the influence of study features in their overall aggregate analyses, and they may not devote as much of their inquiry to moderators. For other types of reviews, the goal to understand and model variability in the study findings may call for wider inclusion/exclusion criteria and an extensive focus on moderator analyses to understand how generalizable the results might be.

It is essential to provide a rationale for the moderators that are selected before the meta-analysis is conducted, to avoid "fishing" for results by trying out multiple moderators until one is statistically significant – a very bad idea in any statistical analysis. Coding the moderators for each study is one of the challenges in moderator analyses. Inevitably, relevant information about moderators is often not provided in many primary research reports, resulting in only a subset of the studies being suitable to be included in these analyses.

Accounting for Moderators: Subgroup Analysis and Meta-Regression

In a subgroup analysis, a categorical variable splits the mean effect size into group-specific effect sizes, which are then compared statistically. For example, one might investigate how the response to a particular pedagogic strategy differs between students in honors vs. regular classes or in small vs. large classes. A statistically significant subgroup analysis signifies that this additional variable does play a role in the observed effect. The classification of the groups to test should be identified based on a scientific hypothesis before starting the analysis, and the means and confidence intervals of each group are usually presented.

Meta-regression is an extension of regression in primary data analysis but includes weighting by both sampling variance and random effects variance components. It is a powerful advance over the subgroup analysis approach because multiple covariates (moderators, predictors, or explanatory variables) and both continuous and categorical covariates can be included in a single statistical model. One could ask if family income, grade point average, and membership in an honors class influence the relationship between the experimental education program and knowledge retention, when all are considered together. It is rare (and can be problematic) to include more than a few covariates in a meta-regression model. The choice of covariates should be pre-specified and informed by sound theoretical foundations. The more covariates, the more difficult to interpret the results can become. As a result of the power and versatility of meta-regression, it is nevertheless often the preferred method for examining the impact of covariates.

In a meta-regression, the (weighted) linear regression results are presented graphically, and in all cases the statistical results will be reported. Each regression coefficient is accompanied by a probability that the observed effect is due to chance. Approaches have been recently introduced for determining the best statistical meta-regression model where several moderators are being tested (see, e.g., Cinar et al., 2021).

Interpreting Meta-Analytic Results

Exploratory Data Analysis and Display

Understanding the data structure is critical. Graphical tools can help you to better understand whether the assumptions of the analysis are met, if there is confounding or unbalance in the moderators you want to test, whether various kinds of biases exist, and what the statistical distributions of the data are. Histograms, weighted histograms, normal quantile plots, and other graphical tools can be used to understand the distribution of the data. Simple contingency tables listing the number of studies with particular characteristics can be very helpful in identifying imbalance and confounding (Table 27.1). Funnel plots graphically depict the relationship between studies' effect sizes and sample sizes and can be used to signal publication bias.

As mentioned above, the results of a meta-analysis are almost always presented as a forest plot (see Figure 27.1), which visually represents the effect size magnitude and precision for individual studies and the overall effect size and precision for the pooled studies. In a *cumulative meta-analysis* plot (Figure 27.2) each effect size is

Table 27.1 *Contingency table for numbers of studies (e.g., loss in automobile value over time for used cars) with values indicated for two moderators: color and size**

Color Height	Red	Green	Blue	Purple
Small	6	0	0	2
Mid-size	11	0	0	0
Large	2	12	11	0

* One cannot say anything about the effect of color on resale value in mid-size cars; in comparing the effects of size one mostly has information only about red automobiles.

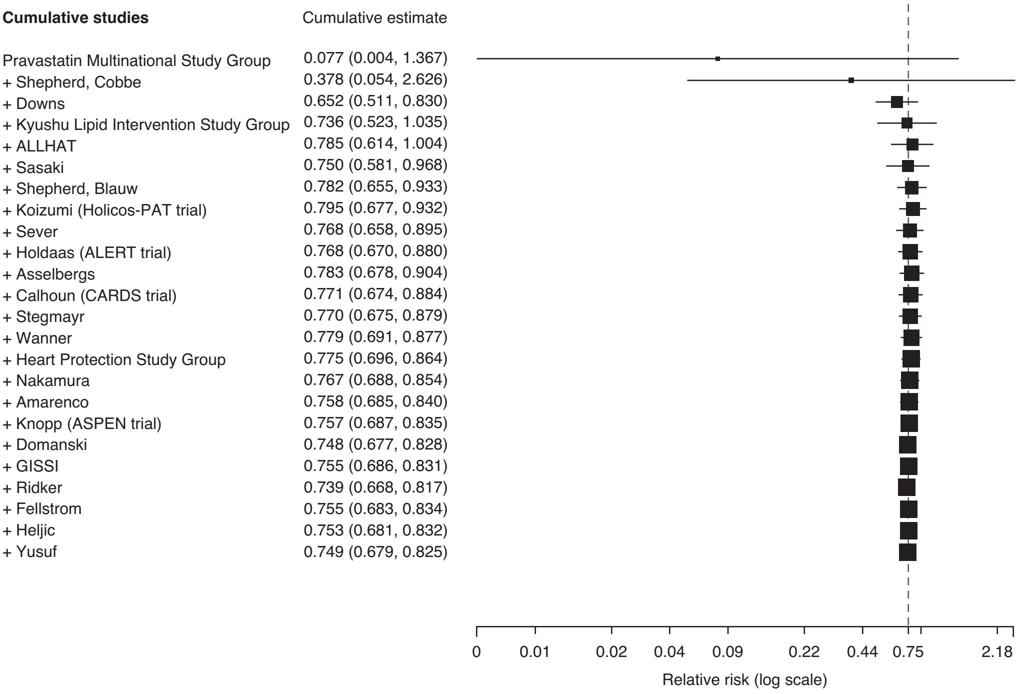


Figure 27.2 *Example of a cumulative meta-analysis, showing the evolution of the mean effect size over time.*

a recalculation of the pooled effect size, with each study added to the pooled result, one at a time in chronological order. The last effect size at the bottom of the graph is the same as the pooled effect size in a traditional forest plot. This type of plot is helpful in understanding the evolution of a body of evidence over time and can help identify the point at which there was sufficient primary research data available to reach definitive conclusions.

Unbalanced Data

Unlike the planned structure in a primary study (e.g., a randomized control trial or other experimental design), meta-analytic data are not usually planned. This can create severely unbalanced data sets, which present logistical challenges for analyses. Missing data for moderators in some studies means that a complete statistical model cannot be tested on the entire data set. There may be many studies on some aspects of a question and little information on others. For example, perhaps the meta-analyst is interested in how well a certain intervention, designed to increase memory retention, works on young vs. elderly participants. They find 28 studies on this effect on college students, but only two studies on senior citizens. This is an unbalanced data set because there is a lot of information for one group and little for the other group. It would therefore be difficult to get a convincing answer to the question of whether the method works equally well for both populations. There is unfortunately no easy solution to this problem, but the general principles of meta-analysis still apply, including assuring that a systematic review was conducted so that all available studies are included, and carefully reporting results to acknowledge when major limitations are present. One of the most valuable contributions of the meta-analysis can be to point to where more research is needed (e.g., on senior citizens) and where enough information is already available; it is a waste of time and funding to keep doing more research when the outcome is already well established and understood.

Confounding

Confounding occurs when two or more moderators are closely related, or inseparable statistically, threatening the interpretation of the results. Perhaps you want to determine how large and small hospitals differ in the effects of a certain intervention to prevent post-surgical re-admission and whether the effect differs in rural vs. urban settings. However, all the large hospitals are in cities, and all the small hospitals are in rural areas. A test of the effect for large vs. small hospitals cannot distinguish between that effect and the effect in rural vs. urban settings. The only solution is greater awareness and care in making inferences from such a meta-analysis. One can say that large urban hospitals have a different effect from small rural hospitals, but that’s about it.

Publication Bias, Research Bias, and Methodological Bias

The validity of any review or synthesis of research, whether it is a meta-analysis or narrative review, is compromised by the omission of relevant data. Publication bias, whereby research studies do not make their way into the published literature where they can be readily identified, is a well-established threat to the validity of research syntheses. Distortion of sound conclusions may occur when particular types of studies are more likely to be found in the published literature (e.g., those that find significant effects or effects that are in the expected direction), while others languish unpublished in file drawers (i.e., the “file-drawer problem”).

Conclusions based on biased literature are inaccurate and statistically biased when only particular kinds of outcomes are reported. This type of preferential selection can occur at the level of journal editors, reviewers, or investigators themselves. One illustrative example comes from an investigator’s own work, reporting on the discrepancy in the fates of findings that were published versus remaining unpublished (Lane et al., 2016). Across 8 studies on the effects of administering the hormone oxytocin intranasally on social and emotional behavior, from his own laboratory assessing 13 different dependent variables, only 5 papers were accepted for publication (a 38% publication rate). Studies with null findings or that obtained findings that were not in the expected direction were more likely to be rejected for publication. When these authors meta-analyzed their entire published and unpublished portfolio of studies on this topic, they found that the aggregated effects of intranasal oxytocin were not reliably different from zero.

A related issue occurs when a widely cited seminal paper that produces intuitively appealing findings for a scientifically intriguing question cannot be replicated later, or when studies that fail to confirm the results of such publications remain unpublished or are not even conducted. Publication bias has been demonstrated in many different fields (e.g., Cassey et al., 2004; Dickersin, 1990; Forstmeier et al., 2017). A vigorous push for “open science,” emphasizing transparency, repeatability, and full disclosure of data, has been acting to counter this problem (Nosek et al., 2015).

Procedures that can suggest the “fingerprints” of publication bias include the funnel plot, which highlights gaps where small studies with small effects may have gone unpublished, and formal indices (e.g., Egger’s test; Egger et al., 1997). Rosenthal’s fail-safe *N* (Rosenthal, 1979) and refinements (Orwin, 1983; Rosenberg, 2005) may be useful exploratory tools but have limitations. More rigorous tests of publication bias can be carried out by comparing unpublished results (e.g., “gray literature”) with the results of published studies. Ultimately, the only solution for publication bias is a change in perception and practice, whereby the statistical significance of a study does not determine if it is published. Considerable efforts to change publication practices have been motivated by this realization (Ioannidis, 2018).

Perhaps even more pervasive than publication bias is what has been called research bias (Gurevitch & Hedges, 1999). Research bias occurs when only certain things are studied and reported in the scientific literature; it is ubiquitous but generally under-recognized. If research bias exists, any summary of the literature will omit those things that have not been studied or are under-studied, giving a biased view of reality. For example, medical studies have often historically focused only on men and excluded women (Hamberg, 2008). Summaries of the results of those studies may have provided information on how men responded to various medical treatments, but physicians are left guessing about women's responses. Similarly, many psychological studies are carried out on limited or non-representative populations, including populations only from Western, educated, industrialized, rich, and democratic (i.e., WEIRD) countries or consisting of convenient-to-study college students; their results may not be applicable to the responses of people from other populations (e.g., children, older adults, or those from other countries; Henrich et al., 2010).

In ecology and many areas of molecular biology, certain organisms and biological systems are well studied, but the literature may tell us little about those that are ignored. For example, Lowry et al. (2013) found that biological invasions were well studied in North America, Europe, and several other geographic areas, but largely ignored in tropical regions, presenting a misleading overall picture of where biological invasions are prevalent or problematic. Far more research has been conducted on cystic fibrosis, which affects predominately White children, than on sickle cell disease, which predominately affects Black children; this research bias results in limited understanding of appropriate treatments for those affected by sickle cell disease (Farooq et al., 2020). Research bias is profound, far-reaching, and can undermine any meta-analysis.

"Not Significant" Main or Moderating Effects

Sometimes, one finds that either the mean overall effect or the moderators of interest are not statistically significant; this may seem disappointing. There are several reasons that the analysis may yield results that are not statistically significant, and it is important to present one's findings regardless of whether they are "significant" because this is distinct from "meaningful" or "biologically or clinically important." There may really be no effect – the new wonder drug may truly be no different in its outcome than the cheap old one; although it may be tempting not to report that, it is really valuable information. Hypothesized moderators may, in fact, not change anything; men may respond in the same way as women or socio-economic factors may not influence learning outcomes. You may also not have sufficient statistical power to detect a real effect, even though there really is an effect. There may be limitations in the experimental designs of the studies (e.g., too short a study duration or the ages of the participants may be too limited). Researchers (and readers) should be alert to these and other limitations of meta-analyses.

Correlation Not Causation: The Nature of Synthesis-Generated Evidence

Synthesis-generated evidence has both advantages and disadvantages over evidence from individually designed experiments. Because a meta-analysis summarizes the available evidence, non-existing information will not be included in the synthesis. If all the studies are short-term, the meta-analysis cannot provide information on long-term responses. If biases in the literature exist such that particular populations are unstudied, the summary cannot be generalized to include those missing populations and missing information.

Controversy exists regarding the extent to which the evidence from a meta-analysis can be the basis for testing hypotheses. On the one hand, it summarizes the available evidence. That is the basis for using meta-analysis results in evidence-based medicine and other evidence-based decisions. On the other hand, evidence may still be merely correlational. Importantly, however, such synthesis-generated evidence can address questions that have not been directly investigated in primary studies. This is enormously important in ecological research, in which experiments are typically carried out at small spatial scales in particular locations and where meta-analysis can provide regional and global syntheses across many studies. Large-scale, more general pictures of responses can also be valuable in social and behavioral science meta-analyses. A meta-analysis might include people from a broad range of geographic areas and ethnicities, or it may provide comparisons between long-term effects and studies of shorter duration. Research syntheses can also identify research gaps in areas for additional research or theory development.

Advanced, New and Emerging Evidence Synthesis Methods

The field of meta-analysis continues to grow and evolve to accommodate a greater diversity of evidence, methods, and decision-making stakeholders. Qualitative evidence is now widely accepted as a form of valid research-derived data that can inform practice and policy, while diagnostic test accuracy and health economic evaluation evidence are found to be increasingly valuable to guide medical decision making and resource allocation, respectively. Here, we review some of these newer tools used in research syntheses, their core principles, and current or potential applications.

Qualitative Evidence Synthesis

The advent of qualitative evidence as being viewed as a legitimate scientific endeavor has led to an explosion of qualitative research studies being published; with them, the same problem emerged of more data than individual decision makers can ever keep up with. To address this challenge, new methods for qualitative systematic review have been put forth and, in some cases, used extensively (Lockwood et al., 2015). The steps of a qualitative systematic review are similar to those for a quantitative systematic review, with the exception that critical appraisal or research quality determination for qualitative evidence consists of the assessment of rigor for methodologies that are completely different from quantitative research; in particular, the data synthesis piece is truly very different. The first crucial point is to recognize that, in qualitative research, the data to be pooled are no longer numerical summary measures – they are text, or "textual data" to be more exact, which can consist of words on a page, video or audio, or images (Lockwood et al., 2015). The pooling of this textual data is informed by a variety of qualitative research methodologies (e.g., realist synthesis, aggregative synthesis, or ethnographic synthesis), but they all share one feature – they combine themes (which summarize large swaths of textual data at the primary research level) from multiple qualitative studies and present a larger body of these themes that can be used to better inform decision making. The differences between them rest in how they combine those themes and the relative level of selectivity or inclusivity of each of these approaches regarding the primary research study themes.

Network Meta-Analysis and the Meta-Analysis of Networks

In a pairwise meta-analysis, there are two variables being compared: one predictor variable (also known as the independent variable in the experimental research arena) and one outcome variable (i.e., the dependent variable). Regardless of how many groups the predictor variable contains, only two of those groups can be compared within any given simple meta-analysis. For example, in a meta-analysis comparing the effectiveness of three different drugs on a given outcome (the independent variable "drug received [A or B or C]" has three groups), a simple pairwise meta-analysis can compare drug A vs. drug B, drug B vs. drug C, or drug A vs. drug C. The closest equivalent of this type of simple meta-analysis in primary research is bivariate statistics with a dichotomous predictor (i.e., a variable that has only two groups). In this simple meta-analysis scenario, outcome data from studies where the predictor is dichotomous are extracted and pooled. As in primary research, this outcome data will be pooled as a mean difference (if the outcome is continuous) or as a risk ratio or odds ratio (if the outcome is dichotomous). The solution to this limitation is the development of network meta-analysis.

Network meta-analysis has at its foundation simple pairwise meta-analysis. Also known as multiple comparison meta-analysis, the novelty of network meta-analysis is that it goes beyond the pairwise comparison, and in fact it has no inherent limit as to how many comparisons it can make at the same time. For example, suppose there are 12 different pedagogic tools that can be used by educators to help improve knowledge retention among high-school students. Suppose that these 12 tools have all been studied extensively in primary research (i.e., there are dozens of studies on them), but in all these studies, each tool has only been compared to a "control" (e.g., the standard pedagogic tool commonly used). Let's say there is now a need to synthesize the literature on these tools to make a funding decision, but it is not clear from the dozens of studies which tools are more effective because all the tools appear to be more effective than the control. How do you determine the tools' effectiveness relative to each other, quickly and efficiently, without needing to spend additional research resources and dollars to test these tools against each other in new experimental studies? Enter network meta-analysis. Figure 27.3 provides a graphical example of what network meta-analysis accomplishes.

By knowing the effect sizes from the pairwise comparison of each intervention relative to the conventional control approach, it is possible to then indirectly estimate the effect size that would result if a study was conducted that compared two of the interventions of interest directly. If I know the effect of intervention A relative to intervention B, and I know the effect of intervention B relative to intervention C (via primary research studies), then I can use network meta-analysis to combine

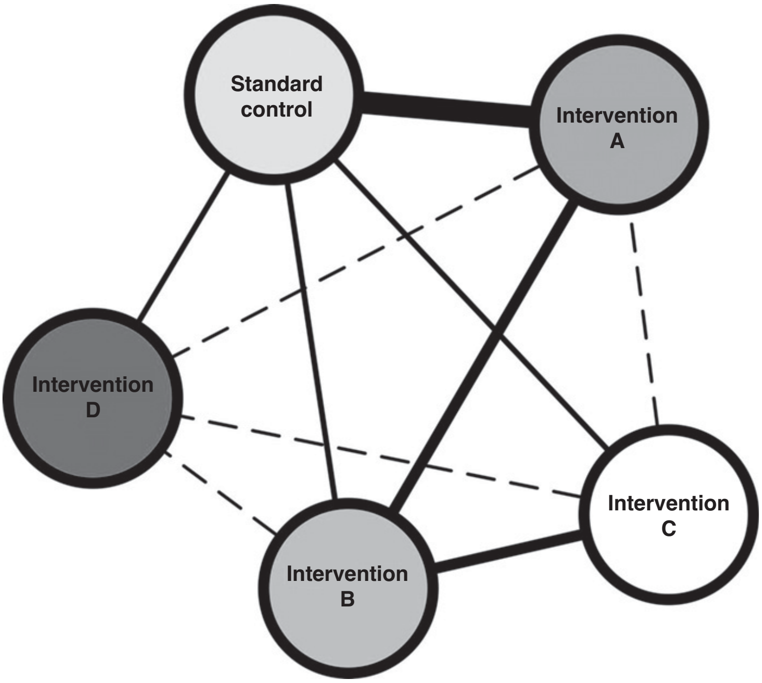


Figure 27.3 *Network meta-analysis graphical display. Thicker lines signify more included studies. Solid lines indicate direct comparison of the interventions within primary studies, and dotted lines represent indirect comparisons derived from the network meta-analysis.*

those primary research study effect sizes and indirectly calculate the effect of intervention A relative to intervention C without needing to conduct a primary research study directly comparing A to C.

Not only is it possible to carry out network meta-analysis as above, but meta-analysis of networks in primary studies can also be carried out. Kinlock (2019) carried out a meta-analysis of the network structure of 31 ecological studies to determine the generality of competition, winner–loser relationships, and unequal interaction allocation in plant communities. She synthesized networks of competitive and facilitative interactions and developed new methods to quantify variation in network structural metrics among studies. This approach may prove to be valuable in other systems in which a synthesis of networks promises to reveal more general patterns.

Diagnostic Test Accuracy Meta-Analysis

Diagnostic test accuracy studies are a type of primary research design that seeks to establish the quality of newer tests (e.g., biochemical laboratory tests, cell cultures, or a new screening tool) relative to gold-standard or reference tests. These types of studies are essential for everyday decision making because they establish how well these tests can confidently detect diseases or conditions of interest. The principal outcomes of these studies are sensitivity (i.e., how likely you are to test positive if you have a particular disease or condition of interest) and specificity (i.e., how likely you are to test negative if you do not have a particular disease or condition of interest); together, they provide a measure of how good a new test is. The application of methods for the systematic review and meta-analysis of these studies (Campbell et al., 2015) have gained popularity in a variety of disciplines, including medicine, nursing, and education.

Imputation, Model Selection, and Machine Learning

Frequently, data on either moderators of interest or on information needed to calculate effect sizes are missing from some of the studies that would otherwise be useful to include in a meta-analysis. In the past, these studies might have been eliminated from the research synthesis, but that could remove useful information. An alternative is to use methods to essentially estimate what these values might be. These various approaches are known as imputation (Kambach et al., 2020). Despite limitations and assumptions, they can provide a tool for using valuable information that would otherwise be difficult or impossible to include in a meta-analysis.

In meta-regression, sometimes several different statistical models are tested; that is, different combinations of covariates may be investigated. One approach to choosing which one best "explains" the data is to use Akike information criteria (AIC) and related tests that balance explaining the greatest amount of heterogeneity while using the fewest possible covariates (Cinar et al., 2021). Meta-regression can be used to examine the association between effect size estimates and the characteristics of the studies included in a meta-analysis, using regression-type methods. By searching for those characteristics (i.e., moderators) that are related to the effect sizes, we seek to identify a model that represents the best approximation to the underlying data-generating mechanism. Model selection via testing, either through a series of univariate models or a model including all moderators, is the most commonly used approach for this purpose. Other approaches involve model averaging. This is a complex advanced topic, which we mention for those who want to pursue it further (see Hobbs & Hilborn, 2006 for an introduction to this literature).

Machine learning, or the use of artificial intelligence (AI) approaches to accomplish previously human-driven functions, has been introduced in several fields for systematic reviewing because of the increasingly large size of the literature (Marshall & Wallace, 2019). Examples of the functions increasingly being performed by AI-driven approaches include searching the literature, screening studies by title and abstract, critical appraisal, and data extraction. Other uses of machine-learning algorithms and other computational techniques have also been developed for meta-analysis, but they must always be guided by a meta-analyst who is knowledgeable both about the subject matter and the methodology of meta-analysis.

Software Programs for Meta-Analysis

Using ordinary software designed for regression, generalized linear models, and other statistical methods for primary data will generally not give you the correct analyses for a meta-analysis or meta-regression. The software you use for carrying out a meta-analysis with or without meta-regression must be able to do several things and may be able to do many others. It must provide all of the information you will need to interpret and report your meta-analysis, including calculating properly weighted effect size measures, the sampling variances, and confidence intervals for these effect sizes for individual studies; importantly, this includes the random effects variance (τ^2) with options for how that is calculated (Boedeker & Henson, 2020). It should calculate the mean effect across all studies (sometimes called the grand mean), its confidence interval, and report heterogeneity statistics. It should provide graphical options for forest plots, funnel plots, and other exploratory graphical tools. The software must be able to calculate weighting based on sampling variances for fixed effect models; for random effects models, weighting must be calculated based on sampling variances plus random effects variances. This capability is also essential for calculating the heterogeneity statistics that are necessary for a meta-analysis.

Some software is open access or inexpensive, yet some is costly. The packages also differ greatly in their ease of use and flexibility (oddly enough, this has no relationship to cost). There are many other useful tools that meta-analysis software can provide, including the capability for carrying out subgroup analyses and more complex meta-regressions, randomization approaches as an alternative to parametric estimation (Adams et al., 1997), tests for model fit (e.g., AIC), more extensive options for exploratory data analysis and graphics, and clearly understandable error messages. Table 27.2 provides a list of some of the software packages for meta-analysis and some of their characteristics. One of the most notable packages is the R package metafor. This package is highly comprehensive, regularly updated, open access, but somewhat challenging to use and interpret, particularly for categorical covariates (moderators). Drawbacks to such a complex and comprehensive package are the steep learning curve and the ease of doing analyses incorrectly without realizing it. Packages also exist for data management and the search and selection tools important for the complete process of systematic reviews. Examples of this diverse group of packages include MetaGear (Lajeunesse, 2016), RevMan (*RevMan* 5.4 2021), JBI SUMARI (Munn et al., 2019), and others.

Best Practices and Quality Control for Meta-Analysis

Developing Transparent Protocols

As with recent practices in carrying out primary research in the social and behavioral sciences, systematic reviews and meta-analyses are increasingly required to be preregistered to be accepted for publication in peer-reviewed journals. One such

Published online by Cambridge University Press

Table 27.2 *Software packages commonly used to carry out meta-analyses**

Package	Ease of use	Cost	Website	Regularly updated	Full capability
Comprehensive Meta-analysis	High	High	www.meta-analysis.com	Yes	Yes
OpenMEE	High	Open access	www.cebm.brown.edu/openmee	No	Yes
Open Meta Analyst	High	Open access	www.cebm.brown.edu/openmeta	Yes	No
Metafor	Low	Open access	https://metafor-project.org/doku.php	Yes	Yes
RevMan	High	Open access	https://community.cochrane.org/help/tools-and-software/revman-5	Yes	No
Stata	Moderate	High	www.stata.com	Yes	No
JBI SUMARI	High	Moderate	https://sumari.jbi.global	Yes	No

*See also https://jlpm.amegroups.com/article/download/5034/pdf and Schmid et al., 2013). Stata is a general-purpose statistics package with a meta-analysis component; others are dedicated for meta-analysis (some with systematic review capability – RevMan and JBI SUMARI). Full capability indicates that the software provides a wide range of effect size indices, is adapted for use in multiple disciplines, and includes full statistical analysis such as meta-regression and other capabilities.

repository of systematic review protocols is the PROSPERO database (www.crd.york.ac.uk/prospero). Prospectively registering a protocol at inception helps to avoid duplication and prevents reporting bias by providing a permanent record of the procedures that were planned, which can be readily compared with a completed review. Organizing and planning one's procedures in advance can help one think clearly about one's review and have a good plan once a systematic review is initiated. The PROSPERO guidelines suggest that registration should occur before screening reports for eligibility; however, reviews where data extraction from reports has not yet commenced will be accepted.

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines provide a checklist of elements to report (e.g., specifying the sources used to identify studies and the methods used to assess risk of bias) and a flow diagram that documents the steps in the process of including and excluding studies in the review (www.prisma-statement.org). Reporting standards may be somewhat different in different disciplines (e.g., O'Dea et al., 2021 for ecological meta-analyses). Importantly, adhering to the checklist ensures reporting quality but not necessarily methodological quality. Completing the checklist only involves reporting whether a particular rigorous element was or was not conducted. Nonetheless, it provides a list of practices that are considered methodologically rigorous, which can be used as a guide before embarking on a review. The PRISMA checklist is primarily intended for reporting of systematic reviews and meta-analyses of interventions but includes variations for other types of study designs.

Repeatability

Developing procedures that ensure repeatability and reliability in carrying out and reporting your results is important. These procedures typically involve delineation of search details, clear criteria for inclusion and exclusion, a manual that guides raters in making coding decisions, and conducting and reconciling independent ratings made by more than one coder. Independent coders must be thoroughly trained and should meet regularly to prevent coding drift. For the proportion of studies that are double coded (that is, two or more researchers carry out the data extraction from each paper; typically, at least 10% of the reports), inter-rater agreement is calculated and optimally is above 80%. Creating a manual is essential in providing guidance when there are conundrums for which there is no one reasonable answer; for example, if one is coding the number of research participants per group, and all that is reported is the total number of participants, is it reasonable to assume that the groups are of equal size to obtain this critical piece of information? Often a reviewer faces many such judgment calls in the process of coding studies; codifying how to handle these situations for the purposes of transparency and repeatability is invaluable. Thoroughly documenting procedures means that systematic reviews and meta-analyses can be scrutinized and replicated in the same way as primary studies can.

Avoiding Outdated Practices

As the expertise of practitioners and understanding of meta-analytic procedures has evolved, certain more commonly seen practices have fallen out of favor. One egregious example is vote counting, whereby studies that find a statistically significant effect are considered supportive evidence and those that do not find a statistically significant effect are considered unsupportive evidence – they are tallied to come to a conclusion. This approach is biased and statistically flawed, and the results of vote counts are unreliable. Vote counts ignore power and the magnitude (strength) of the effect and provide inaccurate comparisons of the effects of moderators. Other outdated practices are using the results of heterogeneity tests to determine whether a fixed effect or a random effects approach to synthesizing results across effect sizes is appropriate. This decision instead should be based on the scientific understanding about the population of studies that one wishes to generalize to; most often a random effects model is the best approach (Hedges & Vevea, 1998).

Correcting for Measurement Error and Ensuring Conceptual Comparability

Particularly in the social and behavioral sciences, it has often been the practice to correct for measurement errors and artifacts in primary studies. Such corrections can provide more accurate estimates of effect sizes where differences across studies may be due to various inaccuracies of measurement or methodological problems (e.g., lack of replicability among study coders; Card, 2016). Similarly, imperfect validity, the extent to which a measure captures what it is supposed to assess (e.g., depression) can, in some cases, be corrected for (Salgado & Moscoso, 2019).

One common criticism of meta-analysis is the concern about combining findings from dissimilar studies (i.e., "apples and oranges") to draw meaningful generalizations. However, as described above, the tools of meta-analysis allow one to examine variation in important study features that help explain differences in intervention effects across studies. Meta-analysts should use both scientific and statistical judgment, based on one's content expertise in the area under study, to synthesize only studies that could be considered conceptually comparable. Another concern surrounds injudiciously combining studies of varying methodological quality – the "garbage in, garbage out" problem. Aspects of methodological quality can be used either as eligibility criteria to restrict studies to only those of high quality or as moderators to model the relationship between methodological quality and effect sizes. Nevertheless, even lower-quality studies may provide important information, and it is still important to be mindful of the extent to which methodological features, particularly those that were not or could not be examined as moderator variables, might account for some of the observed variation in effect sizes. One should read a meta-analytic study with the same type of critical consumer stance as one would read a report of a primary study.

Conclusion: Our Past, Your Present, and the Future

Meta-analysis was first developed in psychology in the 1970s, and soon afterwards was applied to medical research and to other social and behavioral sciences. In the early 1990s, it was introduced to ecology, where it has become an important approach. Since then, meta-analysis application and methodology have developed considerably and spread to many other disciplines. Practitioners in the different disciplines have learned from advances in other disciplines, to some extent, and the basic methodology is fairly consistent. However, practices differ in both technical aspects and philosophical approaches among disciplines. For example, medical meta-analyses tend to be far narrower, with far fewer studies and a much more specific focus, while ecological meta-analyses are almost always very expansive in scope and in the questions addressed (Gurevitch et al., 2018). "Thinking structurally" about how research questions in meta-analysis are framed and addressed in other disciplines can be an excellent tool for making advances in meta-analysis in one's own discipline, and extending the tools and methodological and conceptual advances from other disciplines may lead to new insights; it certainly has for the three authors of this chapter, who come from backgrounds in medicine, psychology, and ecology.

For those inspired to embark on a meta-analysis of their own, especially those early in their careers, here are some final words of wisdom. First, if the number of relevant studies is relatively small and the number of variables to be coded is limited, this endeavor does not necessarily require external funding because the material resources needed are minimal. One does require access to bibliographic search engines and statistical software as well as coding time. Also, data collection does not involve recruiting or compensating research participants or obtaining human or animal subject approval. Meta-analyses are also often well cited. Finally, conducting a systematic review with meta-analysis can help one build a deeper understanding of a research area and make an important contribution to the literature.

As research findings continue to accumulate, the need to skillfully summarize and make sense of them will only increase. As the emphasis on evidence-based interventions grows, clinicians, patients, citizens, policy makers, and researchers will be ever more reliant on the guidance of those who can synthesize complex literatures using meta-analytic skills. Meta-analyses have also proliferated, necessitating meta-syntheses and umbrella reviews to make sense of these reviews. Organizations devoted to supporting the conduct and dissemination of up-to-date systematic reviews, meta-analyses, and umbrella reviews, include the Cochrane Collaboration (www.cochrane.org) and Joanna Briggs Institute (JBI) (https://jbi.global), which focus on health interventions, and the Campbell Collaboration (www.campbellcollaboration.org/about-campbell/history.html), which focuses on social interventions. Rapid “living” systematic reviews, which are completed quickly and updated regularly, provide timely evidence and have been deemed particularly useful with the continuously emerging and important health information emerging during the coronavirus (COVID-19) pandemic. We encourage the readers to use this chapter as a starting point for their journey into meta-analysis but to consider exploring these and other resources to keep up with this ever-evolving field.

References

Adams, D. C., Gurevitch, J., & Rosenberg, M. S. (1997). Resampling tests for meta-analysis of ecological data. *Ecology*, *78*(4), 1277–1283. https://doi.org/10.1890/0012-9658(1997)078[1277:RTFMAO]2.0.CO;2

Boedeker, P. & Henson, R. K. (2020). Evaluation of heterogeneity and heterogeneity interval estimators in random-effects meta-analysis of the standardized mean difference in education and psychology. *Psychological Methods*, *25*(3), 346–364. https://doi.org/10.1037/met0000241

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. John Wiley & Sons.

Campbell, J. M., Klugar, M., Ding, S., Carmody, D. P., Hakonsen, S. J., Jadotte, Y. T., White, S., & Munn, Z. (2015). Diagnostic test accuracy: methods for systematic review and meta-analysis. *International Journal of Evidence-Based Healthcare*, *13*(3), 154–162. https://doi.org/10.1097/xeb.0000000000000061

Card, N. A. (2016). *Applied Meta-Analysis for Social Science Research* (paperback edition). Guilford Press.

Cassey, P., Ewen, J. G., Blackburn, T. M., & Møller, A. P. (2004). A survey of publication bias within evolutionary ecology. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *271* (suppl 6), S451–S454.

Cinar, O., Umbanhowar, J., Hoeksema, J. D., & Viechtbauer, W. (2021). Using information-theoretic approaches for model selection in meta-analysis. *Research Synthesis Methods*, *12*(4), 537–556. https://doi.org/10.1002/jrsm.1489

Dickersin, K. (1990). The existence of publication bias and risk factors for its occurrence. *JAMA*, *263*(10), 1385–1389.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*(7109), 629–634.

Farooq, F., Mogayzel, P. J., Lanzkron, S., Haywood, C., & Strouse, J. J. (2020). Comparison of US federal and foundation funding of research for sickle cell disease and cystic fibrosis and factors associated with research productivity. *JAMA Network Open*, *3*(3), e201737. https://doi.org/10.1001/jamanetworkopen.2020.1737

Forstmeier, W., Wagenmakers, E., & Parker, T. H. (2017). Detecting and avoiding likely false-positive findings–a practical guide. *Biological Reviews*, *92*(4), 1941–1968.

Gurevitch, J. & Hedges, L. V. (1999). Statistical issues in ecological meta-analyses. *Ecology*, *80*(4), 1142–1149. https://doi.org/10.1890/0012-9658(1999)080[1142:SIIEMA]2.0.CO;2

Gurevitch, J., Koricheva, J., Nakagawa, S., & Stewart, G. (2018). Meta-analysis and the science of research synthesis. *Nature*, *555*(7695), 175–182. https://doi.org/10.1038/nature25753

Hamberg, K. (2008). Gender bias in medicine. *Women's Health*, *4*(3), 237–243.

Hedges, L. V. & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Elsevier Science. http://qut.eblib.com.au/patron/FullRecord.aspx?p=1901162

Hedges, L. V. & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*(4), 486–504.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Higgins, J. P. T. & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558.

Hobbs, N. T. & Hilborn, R. (2006). Alternatives to statistical hypothesis testing in ecology: A guide to self teaching. *Ecological Applications*, *16*(1), 5–19. https://doi.org/10.1890/04-0645

Ioannidis, J. P. A. (2018). The proposal to lower *p* value thresholds to .005. *JAMA*, *319*(14), 1429–1430. https://doi.org/10.1001/jama.2018.1536

Kambach, S., Bruelheide, H., Gerstner, K., et al. (2020). Consequences of multiple imputation of missing standard deviations and sample sizes in meta-analysis. *Ecology and Evolution*, *10*(20), 11699–11712. https://doi.org/10.1002/ece3.6806

Kinlock, N. L. (2019). A meta-analysis of plant interaction networks reveals competitive hierarchies as well as facilitation and intransitivity. *American Naturalist*, *194*(5), 640–653. https://doi.org/10.1086/705293

Lajeunesse, M. J. (2016). Facilitating systematic reviews, data extraction and meta-analysis with the metagear package for R. *Methods in Ecology and Evolution*, *7*(3), 323–330. https://doi.org/10.1111/2041-210X.12472

Lane, A., Luminet, O., Nave, G., & Mikolajczak, M. (2016). Is there a publication bias in behavioural intranasal oxytocin research on humans? Opening the file drawer of one laboratory. *Journal of Neuroendocrinology*, *28*(4). https://doi.org/10.1111/jne.12384.

Lockwood, C., Munn, Z., & Porritt, K. (2015). Qualitative research synthesis: Methodological guidance for systematic reviewers utilizing meta-aggregation. *International Journal of Evidence-Based Healthcare*, *13*(3), 179–187.

Lowry, E., Rollinson, E. J., Laybourn, A. J., et al. (2013). Biological invasions: A field synopsis, systematic review, and database of the literature. *Ecology and Evolution*, *3*(1), 182–196. https://doi.org/10.1002/ece3.431

Marshall, I. J. & Wallace, B. C. (2019). Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. *Systematic Reviews*, *8* (1), 1–10.

Munn, Z., Aromataris, E., Tufanaru, C., et al. (2019). The development of software to support multiple systematic review types: The Joanna Briggs Institute System for the Unified Management, Assessment and Review of Information (JBI SUMARI). *International Journal of Evidence-Based Healthcare*, *17*(1), 36–43.

Nosek, B. A., Alter, G., Banks, G. C., et al. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

O'Dea, R. E., Lagisz, M., Jennions, M. D., et al. (2021). Preferred reporting items for systematic reviews and meta-analyses in ecology and evolutionary biology: A PRISMA extension. *Biological Reviews*, *96*(5), 1695–1722. https://doi.org/10.1111/brv.12721

Orwin, R. G. (1983). A fail-safe *N* for effect size in meta-analysis. *Journal of Educational Statistics*, *8*(2), 157–159.

Peters, M. D., Godfrey, C. M., Khalil, H., et al. (2015). Guidance for conducting systematic scoping reviews. *International Journal of Evidence-Based Healthcare*, *13*(3), 141–146. https://doi.org/10.1097/xeb.0000000000000050

Rosenberg, M. S. (2005). The file-drawer problem revisited: A general weighted method for calculating fail-safe numbers in meta-analysis. *Evolution*, *59*(2), 464–468.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638.

Salgado, J. F. & Moscoso, S. (2019). Meta-analysis of the validity of general mental ability for five performance criteria: Hunter and Hunter (1984) revisited. *Frontiers in Psychology*, October 17. https://doi.org/10.3389/fpsyg.2019.02227

Schmid, C., Stewart, G., Rothstein, H., & Lajeunesse, M. (2013). Software for statistical meta-analysis. In J. Gurevitch & K. Mengersen (eds.), *Handbook of Meta-Analysis in Ecology and Evolution* (pp. 174–192). Princeton University Press.

Snilstveit, B., Vojtkova, M., Bhavsar, A., Stevenson, J., & Gaarder, M. (2016). Evidence & gap maps: A tool for promoting evidence informed policy and strategic research agendas. *Journal of Clinical Epidemiology*, *79*, 120–129.

28 Qualitative Analysis

Nicky Hayes

Abstract

In this chapter, we will discuss the "big four" approaches to qualitative analysis – qualitative content analysis, thematic analysis, grounded theory, and discourse analysis – before briefly describing four additional commonly used approaches. Some of these approaches are empirical, either theory-driven or inductive, identifying observable concepts in the data. In others, research is from a social constructionist perspective, incorporating the researcher's interpretation as an essential part of the analysis. Some methods, such as thematic analysis, can be used for either approach. This epistemological range means that, as with quantitative analyses, it is essential to select the appropriate method for analyzing the data, and the rigorous procedures involved in qualitative methodology must be followed meticulously.

Keywords: Thematic Analysis, Grounded Theory, Qualitative Content Analysis, Discourse Analysis, Epistemology

Introduction

Qualitative data have always been part of social and behavioral research – from the 1920s, with Margaret Mead's ethnographic studies of South Sea Islanders (Mead, 1928) and Bartlett's serial reproduction studies of memory (summarized in Bartlett, 1932), to Marie Jahoda's study of the effects of unemployment in Marienthal during the 1930s (Jahoda et al., 1932), and the groundbreaking investigation of the social origins of depression in London housewives by George Brown and Tirril Harris in the 1970s (Brown & Harris, 1978). In fact, our knowledge of clinical neuropsychology was, until very recently, almost entirely derived from qualitative rather than quantitative data; the same applies to many significant insights into mental health.

It is faintly surprising, therefore, that the development of robust and reliable techniques for qualitative analysis, until recently, lagged behind those for quantitative approaches. Fortunately, however, this is no longer the case. As interest grew in the subjective and individual aspects of social and behavioral research, a need to evaluate the hermeneutic aspects of human experience, which could not be achieved by quantitative analysis alone, also became apparent. Consequently, the final two decades of the twentieth century revealed an increasing interest in qualitative analysis that continues to grow.

This chapter outlines the "big four" approaches to qualitative analysis: qualitative content analysis, thematic qualitative analysis, grounded theory, and discourse analysis – including issues of epistemology, procedure, and validity. Qualitative content analysis is an empirical approach primarily concerned with establishing meaningful categories to describe the data. It may be inductive – deriving categories from the data, or directive – exploring how pre-established categories apply. Thematic qualitative analysis can be used as a precursor to other methods or as an analytical approach in its own right. Its flexibility means that it can be used within differing epistemological frameworks, including essentialist, constructionist, and contextualist designs; it can be inductive or theory-driven.

Grounded theory offers a more complex and higher level of analysis, involving an inductive approach to the identification of themes within the data. It requires a rigorous iterative analytical cycle that is only complete once saturation has been reached and no further concept or themes can be identified. The fourth of the "big four" – discourse analysis – focuses on social action (i.e., actions that are performed through discourse). It operates on the premise that human cognitions are flexible and negotiated through discourse, so understanding the nature of that discourse is essential in understanding what human beings do. It can deal with a range of types of data but is all about examining discourse and the social purposes that it serves at social, societal, and individual levels.

Qualitative analysis has a long history in the social and behavioral sciences but has only recently become acknowledged as an acceptably rigorous approach to analyzing data. This is partly because conventional evaluations of validity, such as are typically applied in psychometrics, are inappropriate for most qualitative analyses. Instead, the validity or trustworthiness of the analyses is generally established through three overarching criteria: its *credibility* – whether it represents a truthful account; its *transferability* – how it can be applied in different situations or areas; and its *dependability* – the consistency of its findings (Graneheim & Lundman, 2004). While some forms of qualitative analysis, notably discourse analysis, adopt different criteria, these three are generally regarded as appropriate validation criteria for most types of qualitative analysis.

As researchers have refined older methods and established more rigorous techniques, the old suspicion of qualitative analysis as anecdotal and unreliable has been replaced by increasing respect for the value of qualitative research. Some projects are purely qualitative while others deal only with quantitative data; many modern projects involve "mixed methods," with qualitative analysis augmenting and adding richness to quantitative research methods. Such decisions depend on both the aims and the epistemological assumptions of the researcher.

Preparing Interview Data

Qualitative data can take many forms, ranging from personal accounts of experience, written summaries (e.g., those found in vignettes or comments in questionnaires), to the in-depth approaches of phenomenological interviewing or ethnography. It can even be non-verbal material, such as visual imagery or artwork.

For the most part, however, qualitative studies tend to use interview data as their source material (see Chapter 20 in this volume).

This requires careful data preparation. The first stage is transcription, so that specific passages can be easily and unambiguously identified through their line numbers. However, transcribing an interview involves more than simply writing down the words that people say. There are paralinguistic components to language that must also be included. Some of these will be indicated on the transcription; for example, an interruption is usually indicated by overlapping text, with the second speech starting immediately below the previous utterance at the point where the interruption began. Other aspects of speech are noted using conventional symbols to indicate how the words were actually spoken. Box 28.1 shows some of the most frequently used symbols used for this purpose.

Some material, such as visual imagery or artwork, is less amenable to transcription. For this, the main part of preparation for the analysis consists of immersion through familiarization. The researcher spends significant time exploring the images, both in detail and from a distance, and considering the potential meanings implied by the artist. This can mean adopting tactics (e.g., festooning the workroom with images of the work involved) to ensure constant familiarization and encourage depth of analysis.

The first stage in conducting any qualitative analysis, then, is preparation and familiarization. From that point, the nature of the analysis can vary considerably, depending on the goals of the research and the philosophical or epistemological orientation of the researcher, but many of them involve deriving themes from a data set.

Box 28.1 Common transcription conventions

[] Brackets: overlapping speech indicated by square brackets

= Where there is no interval between the utterances
(.) A short pause
(2.3) A timed pause (generally used for pauses of 0.4 seconds or longer)
: The sound is drawn out – more colons imply a longer sound.
. A tone of voice indicating that the speaker has finished
, A tone of voice indicating that the speaker has not finished
↑ Rising inflection
↓ Falling inflection
! An animated tone of voice
– An interruption or cut off
() Uncertainty from the transcriber
(()) Gestures or extraneous detail, e.g., cough or doorbell, indicated in italics within the brackets
hhh Audible outbreath
.hhh Audible inbreath
< > The part between the arrows is spoken more slowly
> < The part between the arrows is spoken more rapidly
° ° The part between the degree symbols is spoken more quietly
bold Shouting or high-volume talk

Deriving Themes

Theme derivation is what differentiates qualitative research from key examples or illustrative anecdote. The theme is central to the validity of the analysis – its purpose is to represent accurately the subjective meanings and/or social realities that are present in the data (Hesse-Biber & Leavy, 2011). However, meanings can be complex, and identifying all meanings in a report often requires the active involvement of the researcher (Krauss, 2005).

Although there are small variations between methods (Ryan & Bernard, 2003), the essential process of theme derivation is usually quite similar and involves four distinct phases. These are: *initialization* – reading through transcriptions, highlighting what appear to be key phrases or content, coding them, and writing reflective notes; *construction and classification* – producing a labeled initial set of descriptions that may or may not eventually become the themes; *rectification* – reflecting on and stabilizing the themes, and identifying their relationship with established knowledge; and *finalization* – clarifying how the themes relate to the overall storyline of the analysis (Vaismoradi et al., 2016). Theme derivation, then, is at the core of most qualitative analysis. It forms almost the whole of a thematic content analysis, is at the heart of the analysis in thematic qualitative analysis, and is the essential initial component of grounded theory.

Content Analysis

Content analysis is probably the first and, in its simplest form the most straightforward, of all the methods used in qualitative research. It has a long history and has been used for both qualitative and quantitative analysis since the early twentieth century in the USA and the UK (earlier than that in Scandinavian research; Hseih & Shannon, 2005). However, the growing dominance of quantitative analysis through the twentieth century meant that content analysis became regarded purely as a quantitative method, in which data are categorized, counted, and then subjected to other forms of quantitative analysis. More recently, however, its potential as a qualitative method has been increasingly recognized, particularly in applied research.

The important distinction between quantitative and qualitative content analysis is that, while the former is essentially about counting (either words or categories), the latter is about identifying categories with similar meanings, to generate a deeper understanding of what is being described (e.g., Downe-Wamboldt, 1992; Weber, 1990). As such, approaches to qualitative content research have become increasingly more sophisticated. Hseih and Shannon (2005) distinguish between three types of content analysis – conventional, directive, and summative – which vary according to the purpose of the research and the theoretical orientation of the researcher. Their paper gives examples of how each might be applied to a study of end-of-life care.

Conventional content analysis involves coding data that have been directly obtained through the relevant research material; this often involves interviews or accounts but sometimes also documents or texts. It is generally useful where there are few established theories or only limited research data. In this approach, the initial coding is of key concepts that emerge from the data (e.g., the expressions of emotion evident in interview transcripts in a study of end-of-life care by a researcher comparing perceptions of new and longer-term hospice residents; Hseih & Shannon, 2005). The outcome of such a study of end-of-life care might then be compared with an established theory, such as the Kübler–Ross model.

Directive content analysis is theory-driven, so an end-of-life study following this approach might begin with the Kübler–Ross model, rather than bringing it in at the end (Hseih & Shannon, 2005). The overall purpose of the research is to explore the range or validity of the theory, so the data coding follows a predetermined structure, consisting of key concepts or variables derived from the theory under investigation (Potter & Levine-Donnerstein, 1999). Relevant examples are highlighted in the interview transcripts and coded accordingly. Those codes, with exemplars, are used to amplify the subsequent discussion of theory.

Summative content analysis is essentially about exploring how ideas or themes have been used. It begins with a traditional quantitative content analysis – simply counting the number of times given words or textual content appear. In the end-of-life care example provided by Hseih & Shannon (2005), the text consists of discharge teaching for patients transferring from hospital to hospice; this is compared to clinician discussions with patients or family regarding planning for end-of-life care. The content analysis concerned the use of explicit terms (e.g., "die," "dying," and "death") in comparison to euphemisms (e.g., "passing on" or "going to a better place"). This stage is referred to as a manifest content analysis (Potter & Levine-Donnerstein, 1999) and can be used as the precursor to a more interpretive approach that identifies alternative but equivalent terms and explores the underlying meaning of these in the context of the study. That is a frequent approach in document or textual analysis, showing how words are actually being used (Babbie, 1992).

Whichever type of content analysis is being used, the codes are then sorted into categories and grouped into meaningful clusters that inform the subsequent discussion of the research findings (Patton, 2002), and exemplars are drawn from the analysis. Sometimes the data coding is then revisited, at which point the method begins to have a resemblance to thematic qualitative analysis or grounded theory; we will discuss these methods later in this chapter.

Effectively, then, the difference between the three forms of qualitative content analysis is that conventional content analysis begins with observations, and its codes are derived from the observational data. Directed content analysis begins with theory, and its codes are derived from that theory or relevant research. Summative content analysis begins with key terms, derived either from reviewing the literature or from the researchers' interests. Each of them, however, gives an approach to analyzing qualitative data that is more meaningful than simply counting occurrences.

Qualitative content analysis is sometimes referred to in the literature as thematic content analysis. In that respect, some consider it similar to, or even confuse it with, thematic qualitative analysis. The difference can be summarized as description versus interpretation. The ultimate aim of a content analysis is to describe the patterns in the data; interpretation is not its purpose. However, thematic qualitative analysis is all about interpreting data; the descriptive aspects of the analysis are a means towards that end.

Thematic Qualitative Analysis

Although content analysis may be the oldest, thematic analysis is undoubtedly the foundational method for modern qualitative analysis. It is a way of organizing and describing a data set to identify, analyze, and interpret patterns in those data. Some researchers see it mainly as a tool for other methods, such as grounded theory (e.g., Boyatzis, 1998; Ryan & Bernard, 2000), while others regard it as a research method in its own right (e.g., Braun & Clarke, 2006; Hayes, 1997a). Regardless, it is arguably the most frequently used qualitative method in modern qualitative research.

One reason for the popularity of thematic analysis is its epistemological flexibility. Braun and Clarke (2006) discuss how it can be used within an essentialist or realist framework, exploring research participants' experiences and meanings; within a constructionist framework that involves exploring how events and experiences relate to the range of social discourses; or within a contextualist framework, where it focuses on exploring the dynamic interchange between individual meanings and social contexts. This epistemological versatility, however, does not mean that "anything goes." The value of thematic analysis as a research tool is the way that it offers a systematic and rigorous method for extracting meaning from qualitative data, through a series of clearly articulated stages.

Stages of Thematic Analysis

Before commencing any thematic analysis, it is important that the researcher establishes the epistemological framework within which the research is located, since this has implications both for how the themes will be identified and for decisions about what counts as evidence or valid data for the study (See Chapter 20 in this volume). Assuming this has been established, the first stage is data preparation, as described earlier, and familiarization with the data on the part of the researcher.

The second stage is the development of a coding system and its application throughout the data set. At this stage, the codes may represent potential themes, theoretical concepts, or simply interesting features of the data; however, they need to be consistently applied, and all instances where they have relevance must be noted. These are not yet the final themes – the coding involved here is identifying information that can potentially contribute to the final theme(s); it is part of the process rather than an outcome. In a theory-driven thematic analysis, for example, coding involves identifying material indicative of relevant theoretical concepts. For example, Hayes (1997b) showed how an investigation of social identity processes in small companies operationalized the basic theoretical concepts of categorization, intergroup cohesion, and self-esteem into aspects of company life (e.g., the perceived boundaries between personnel, communication, and company pride). The coded data for the thematic analysis were the attributions made about these matters during interviews with company personnel.

It is not until the third stage that themes begin to emerge. Here, coded data are sorted for conceptual similarity. This generates a set of "protothemes" – ideas about possible or potential themes. Protothemes are not fixed and will develop and change as the analysis continues; some initial categories may be rejected as having little relevance, while some categories may be extended to incorporate others. During this stage, the researcher assigns a provisional name to each prototheme and attempts a written definition, while bearing in mind that both can change in subsequent stages of the analysis. The attempt to define the theme in writing is another, distinct part of the analytical process. It helps to clarify the nature and relevance of each potential theme.

The fourth stage is reviewing the protothemes – systematically revisiting the data for each one separately. This essential stage has a dual purpose: to confirm or disconfirm relevance and to pick up relevant information that might have been missed in the initial coding stage. Although it can be a lengthy and tedious process, it is essential; it allows for the selectivity of human perception in earlier stages, or, if a computerized coding system has been applied, ensures that important but implicit information has also been included (Hayes, 2021).The aim at the end of this stage is to produce a "thematic map" of the data, indicating the major themes and giving an idea of how they may be related to one another (Braun & Clarke, 2006). It is the rigor of this process that distinguishes thematic qualitative analysis from a simple thematic content analysis or selection of anecdotal quotes.

The fifth stage involves another pass through the data set to refine the themes and ensure that they are relevant to the overall aims and purpose of the research. This is the point where an appropriate name for each theme will be established, and the written content descriptions that were begun in stage three are adjusted and clarified. At this point, these begin to resemble the descriptions that will appear in the final report.

The final stage of thematic qualitative analysis is preparing the final report. It has three parts. The first is using the now-established themes to structure the selection of clear, illustrative examples from the data set. It is also conventional to identify an illustration, from the data set, that shows how the analysis has proceeded. The second part is relating the analysis back to the original theoretical context for the research – showing the relevance of each theme to the research questions originally derived from the literature. The third part is using this information to produce the final report.

A full thematic analysis is very different from an anecdotal selection of quotes, or even from a thematic content analysis. The re-examinations of the data in various stages are time-consuming, but they ensure a level of academic rigor that provides confidence in the outcomes, enabling either the exploration and/or amplification of existing theory or the emergence of new insights from the data. It is not, however, an easy option for students, and those considering using this method would do well to consider carefully whether an alternative approach might suit their needs and their academic requirements better.

Grounded Theory

The concept of grounded theory was first introduced by Glaser and Strauss (1967), who described it as a way of discovering theoretical concepts in social research data. They argued that it was a general methodology, rather than a simple research method, since it represented an entirely different way of thinking about data. It involves inductive research undertaken by means of rigorous research procedures that allow ideas or theories to "emerge" from the data, rather than being imposed by the researcher (Grounded Theory Institute, 2008).

The fundamental principle of grounded theory, then, is that it does not begin with a specified theory or idea. While the term has often been misapplied, in a grounded theory study the researcher takes an open-minded approach to the data – being interested in what emerges from that data rather than in confirming or validating a pre-existing concept or theory or answering a specific question. At most, there may be a very general research question, defining the area of investigation (e.g., "How do people feel about climate change?"), but any attempt to anticipate, classify, or predict the nature of responses is strictly avoided.

Procedures of Grounded Theory Analysis

Avoiding assumptions is such a basic principle in grounded theory that it shapes how researchers approach the data. Essentially, everything in the data set is potential material for the analysis – from the first line of the first set of text, line by line, to the end. The initial stage is *open coding*; each small section of text is coded – given identifying symbols (words or numbers) that relate to potential key points in the data. Open coding serves two functions: it ensures that nothing is omitted from the analysis, and it also familiarizes the researcher with the material in a systematic way.

Pidgeon and Henwood (1997) cite an example of an investigation into the individual and organizational reasons behind the failure to identify hazardous waste in the redevelopment of an industrial site. Significant concepts in each paragraph of the interviews are identified and given a tentative label. This is further enhanced by *memoizing*, in which the researcher writes memos – possibly full descriptions but more likely shorthand notes – about each of the concepts that have emerged during the open coding. Glaser (1998) described the memoizing process as the core stage of grounded theory. It is a lengthy process, but it produces a bank of ideas from which the researcher can draw later in the analysis. It also helps the researcher to begin to conceptualize concepts that are emerging from the data. Without this stage, Glaser argued, resulting theory will be superficial and unoriginal, since only easily located surface-level concepts from established knowledge frames are likely to be identified.

The categories resulting from the open coding and initial memoizing are then organized to form a preliminary index of those concepts, categories, or labels that seem appropriate to describe each section of text. The label given to each index entry reflects the meaning being perceived by the researcher, but this is likely to be developed and refined as the analysis proceeds since it is only the first of several iterations. The researcher then revisits the entire data set from the beginning, working systematically through it; this includes applying, adjusting, or adding categories or concepts to the index, and where appropriate reassigning text to other, more relevant concepts. Frequency of occurrence is irrelevant – the range and diversity of how the concept is used is more important. Any given concept typically emerges with several different facets, each describing different manifestation of the concept. The indexing system holds the record of these various facets and where they have emerged.

The data are then subjected to several more iterations of the analytic process until the core concepts become saturated – when explorations of the data are no longer presenting new facets or making new connections. The dynamic interchange – the "flip-flop" between the data and the indexing categories referred to by Pidgeon and Henwood (1997) – means that the index develops as the result of a creative process since the researcher's ideas for categories and codings emerge through repeated and reflexive inspection of and reflection on the data.

When saturation has been achieved, the researcher then moves on to the challenging task of writing a comprehensive definition of each category, drawing on the memos that have been written during each iteration. These categories are then sorted and structured, giving the outline of a meaningful theory. The written descriptions of the concepts and categories, and the structure of their connections, helps to clarify and articulate the emergent theory.

There are three types of outcomes that can result from a grounded theory analysis: taxonomy development, local theoretical reflection, or a fully fledged grounded theory (Pidgeon & Henwood, 2004). Taxonomy development is where the grounded theory process is used to develop a set of categories that can be applied in further research; it is particularly valuable when opening up a new research area. Local theoretical reflection is using the analysis to explore a specific case or situation in depth; the method is particularly useful for highlighting unexpected issues or connections. A complete grounded theory will reach beyond the bounds of the immediate topic or data set, covering enough detail that it can be used as a full explanation in a range of contexts; it is, therefore, relevant for new areas of research and for revisiting existing research topics.

Methodological Developments

No formulation of a theory remains without adjustments as it comes into general use. In terms of "classic" grounded theory, there are three basic approaches: that maintained by Glaser (e.g., Glaser, 1998), a slightly modified approach put forward by Strauss and Corbin (e.g., Strauss & Corbin, 1990), and a further adjusted version by Charmaz (e.g., Charmaz, 2014). While the fundamental processes of conducting a grounded theory analysis are essentially similar, there are some philosophical differences between the three, centering on the acceptance or otherwise of achievable access to an "objective" reality. For Glaser, the researcher is neutral – a vehicle for uncovering existing concepts in the data. Strauss and Corbin see this neutrality as implausible and focus instead on minimizing the unavoidable influence of the researcher to minimize their unconscious contamination of the emergent information. Charmaz, on the other hand, adopts a constructionist perspective, seeing the active involvement of the researcher and the importance of their interpretation as key to the analysis and recognizing the outcome as one of a number of possible versions of reality (Singh & Estefan, 2018).

Another issue concerns the relationship between the analysis and pre-existing research literature. Glaser emphasized that the researcher should avoid the literature or any potentially relevant ideas throughout the analytical process. This also prohibits discussion of the research with colleagues or other interested parties, suggesting that this can both detract from and distort the memoizing process and can influence the sorting of the concepts (Glaser, 1998). In Glaser's model, the literature is only consulted and incorporated into a final report once the analytical process has been completed.

This issue of avoiding the literature until the end of the analysis has, however, proved problematic, particularly in PhD research, where the student is expected to present a review of the relevant literature as part of the initial stages of developing a full research proposal. Thornberg (2012) argued that it is possible to conduct a literature review while still retaining a grounded theory approach, proposing several sensitizing principles that can be applied during the literature review to help avoid "contaminating" the research. Above all, Thornberg argued, the researcher needs to adopt a reflexive approach, treating the literature review as a theoretical sampling of the literature, rather than as an attempt to provide a comprehensive account of prior research.

There are wider debates about grounded theory that reach outside of academia. It has become increasingly popular in applied research – in nursing (Singh & Estefan, 2018), information systems (Urquhart et al., 2010), educational research (Thornberg et al., 2015), marketing (Smith, 2020), and many other contexts. As, perhaps, an inevitable outcome of this popularity, it has developed numerous variations. Some of these are relatively unproblematic (Bryant, 2019; Bryant & Charmaz, 2010). However, others have generated controversy, particularly with respect to the misuse of the term "grounded theory," which has been used to refer to a wide range of research techniques, some of which are actually much more simplistic forms of analysis (Suddaby, 2006).

For example, there is sometimes confusion between grounded theory and qualitative content analysis. There are deep differences between them, including their epistemological origins – content analysis follows the empirical tradition while grounded theory emerged from the hermeneutic tradition in social research. Another difference is in the goals of the research – the purpose of a content analysis is essentially description, but the purpose of grounded theory is to generate a theory (or at least an explicative framework). Once coding categories are established, content analysis involves relatively few revisions; this contrasts with the extensive iterations of the coding process in grounded theory. Evaluation also differs, with content analysis following the trustworthiness criteria for qualitative analysis, in general, while the evaluation of a grounded theory is in its conceptual density and theoretical sensitivity (Cho & Lee, 2014).

Grounded theory, then, is very far from being an easy option in qualitative research. Rather, it involves a number of a highly rigorous and demanding stages and requires a real commitment on the part of the researcher. As a method, it offers a rich and detailed account of the area being explored, assists in understanding the nuances of social living, and is capable of generating a range of unexpected or unanticipated insights.

Discourse Analysis

Discourse analysis is the last of the "big four" approaches to qualitative analysis. The primary focus of discourse analysis is the social acts or actions that are performed though the discourse being analyzed, but that discourse can take many forms. It can be interviews or conversations, but it can also be formal speeches (Daghigh & Rahum, 2020), social media (Masroor et al., 2019), media reports (Atawneh, 2008), photography (Beloff, 1997), or even advertising packaging (Parker, 1994).

Many forms of social and behavioral research, whether quantitative or qualitative, tend to focus on observable behavior or human cognitions. However, Edwards (1997) argued that, in general, they fail to take into account how human actions and cognitions are shaped and reshaped through discourse; in other words, what people do or think is flexible and adaptive, rather than fixed. A realistic understanding of human social behavior, or even human society, according to Edwards, must incorporate investigation of the social discourses that shape human experience.

Rhetorical Themes and Interpretive Repertoires

A basic principle of discourse analysis is that discourse serves a purpose. That purpose may be enhancing understanding, emphasizing, or developing power relationships, expressing concerns, reinforcing affiliation, or any number of other objectives. There are two main features of discourse that help to reveal or illustrate that purpose. The first is the *rhetorical themes* that emerge as the analysis shows how different arguments are presented and meanings are constructed. Such themes are used in discourse to illustrate meaning and may become evident in several ways (e.g., through repetition, enumeration, and other rhetorical devices). However, they are most commonly identified through the use of metaphor. Metaphors, such as "pruning the economy" or providing "a shot in the arm" for an ailing business, reveal different underlying ideologies, and ideological conflicts are often revealed by the use of contrasting metaphors for the same events or issues.

Some forms of discourse analysis concentrate exclusively on this aspect of discourse. The method known as thematic decomposition analysis involves specifically identifying those themes that emerge as stories and patterns within the discourse (e.g., Stenner, 1993; Ussher & Mooney-Somers, 2000). They are then examined explicitly in terms of the language used, on the premise that language is a social action, and constitutes social meanings, relevant to the general purpose of the discourse.

The second feature of discourse concerns the *interpretive repertoires* being used by the participants in the discourse – the particular understandings being generated by the content of the discourse. In a study by Nortio et al. (2016), four very different interpretive repertoires were found among Finnish people discussing immigration: one that emphasized immigrants should behave as, and be treated as, a “polite guest”; one that emphasized the need to protect the society’s mainstream culture, that Nortio et al. referred to as “securing the majority”; a “multiculturalist” repertoire that emphasized the need to accept the multicultural nature of society; and an “individualist” repertoire that rejected stereotyping and argued that individuals and their potential contribution to society must be treated as such, rather than stereotyped.

Any given discourse may contain more than one interpretive repertoire. Sherrard (1997) showed how the same people can use several different interpretive repertoires in a discussion, depending on the social action they are performing. In discussions about aesthetic taste, Sherrard found examples of different interpretive repertoires being used by the same person, depending on whether they were making a new point, countering someone else’s argument, or amplifying something they had already said. Commonly, such discourse may contain inconsistencies, but these too provide useful insights as to the underlying social purposes of the discourse.

The Processes of Discourse Analysis

Discourse analysis, like other qualitative methods, requires a high level of immersion in the data. The transcription of verbal material and the intense and contemplative re-reading and exploration of text or imagery is as much a part of this method as it is for grounded theory. The immersion phase for such an analysis will normally take several days, as it is only during this process that the researcher will become aware of deeper levels of meaning in the information.

Coding the data follows the immersion stage, but it differs from other types of analysis in that the coding categories are not about classifying the material but are determined by the focus of the study. The codes provide the researcher with a way of asking questions and a guide for their analysis. They also explore the different ways that the discourse is achieving its effect. Gill (1996) identified four themes that are likely to shape the coding process. The first concerns the discourse itself and how it is formulated – spoken word, texts, images, and so on. The second is how the discourse is being used as a process of constructing social meaning. The third is the overall purpose of the discourse – why it is happening at all, and what its protagonists aim, consciously or unconsciously, to achieve. The fourth is rhetorical organization – how viewpoints are presented or countered through the discourse.

Once the coding has been completed, the researcher is able to begin to develop their report. While this inevitably involves describing the various repertoires or rhetorical themes that have become apparent in the data, the fundamental task is not to describe the discourse but to identify how its underlying purposes are manifest. Discourse is a social act, and as such it is used to achieve social goals. What is of interest in a discourse analysis is how these have been addressed and/or achieved throughout the discourse.

As we saw in the introductory section of this chapter, validity in qualitative analysis tends to focus on ways of establishing the trustworthiness of the data. However, the flexibility and responsiveness of social discourse results in constant change and adjustment. One consequence of this is that it renders traditional approaches to validity and reliability irrelevant; people do not remain consistent and may adopt different interpretive repertoires even within the same conversation. As a result, the assessment of validity and reliability takes a different form. Potter (1996) identified four ways of exploring validity and reliability in discourse analysis. The first is deviant case analysis – examining cases that are unexpected or challenge the patterns or regularities observed in the rest of the data. The second is emphasizing the validity of the person's individual understanding, rather than assuming or imposing a standard or conventional understanding (e.g., where something usually seen as a compliment is taken as an insult). Since this affects the whole nature of the subsequent discourse, it is important that the recipient's personal understanding is taken as more valid than other interpretations.

Third, Potter also argued that reliability and validity may be recognized through the relationship of an analysis with existing research. Its coherence, in terms of cumulative knowledge – whether previous studies are confirmed or disconfirmed by the existing study – gives an important indication of its validity in it social and epistemological context. The fourth criterion identified by Potter concerns the reflexive nature of discourse analysis, and the dynamic exchange between what is being studied and how that investigation takes place. As a result, the interpretations and evaluations of the study made by readers make an important contribution to judgments about its validity.

Critical Discourse Analysis

Some superficial ways of exploring discourse in applied contexts are referred to as discourse analysis. However, these would often be more accurately named content or thematic analyses. The discourse analysis approach in academic research has its roots in critical theory and deconstructionism and therefore tends to adopt a critical approach to the construction of meaning. This makes it very different from the discourse studies typically used in, for example, marketing research. To distinguish the two, it is often referred to as critical discourse analysis.

Ahmadvand (2011) identified three major approaches in critical discourse analysis. The first is the idea of discourse as social practice, with a particular emphasis on the implicit meanings and the unconscious communication of power relationships and other social pressures within the discourse (Fairclough, 1995). Analysis in this approach has three dimensions: description of the discourse, interpretation of the meanings contained within it, and explanation of how these connect with power relationships in society.

The second approach is a socio-cognitive approach to discourse, exemplified in the work of Van Dijk (e.g., Van Dijk, 2006). This approach argues that there is no direct relationship between discourse and social power or organization; rather, that the link is mediated by personal and social cognition. This approach has been described as a triangle consisting of society, cognition, and discourse. People operate with mental models of society that they use to interpret both events and discourse, forming a micro context, but they also act within a macro context of society, power relationships, and inequality. Discourse is a communicative event including a range of forms of interaction, so it benefits from a multidisciplinary approach, drawing on insights from several disciplines, including psychology, linguistics, and semiotics.

The third approach identified by Ahmadvand is a sociological model. This is also interdisciplinary in its acceptance of insights from other disciplines but sees discourse as essentially a form of social behavior, with a dialectical relationship between the discourse and social action (Wodak, 2001). Analysis in this approach is, therefore, primarily descriptive, exploring the relationship between the two.

Discourse analysis, then, is a way of conducting qualitative analysis that has multidisciplinary roots, drawing on insights from a range of social and behavioral science disciplines, including anthropology, linguistics, sociology, psychology, education, and communication studies. As a research method, it adopts a fundamentally constructivist approach, not easily linked with behavioral or empirical research. It is generally used to investigate how discourse is being managed to achieve various social, interactive, or cognitive purposes. Above all, it explores how discourse shapes social understanding.

Other Forms of Qualitative Analysis

While qualitative content analysis, thematic analysis, grounded theory, and discourse analysis are the "big four" of qualitative analysis, they are far from being the only methods of analyzing qualitative information. In a chapter of this nature, there is not space to deal with each one in depth, so what follows is a brief account of four other frequently encountered types of qualitative analysis.

Conversation Analysis

Conversation analysis has occasionally been confused with discourse analysis, but the two involve very different approaches. While discourse analysis explores the overall social meanings and implications of the discourse – what the discourse is actually doing or intended to achieve, conversation analysis operates at the behavioral level – what is actually said and how the conversational exchange is managed. Rather than looking for wider social implications, the researcher explores how the conversation takes place and the various aspects of that process.

The focus in conversation analysis is on how interactions within the conversation are organized. Ten Have (2007) identified four fundamental aspects of conversation analysis. The first is turn-taking organization: how exchanges between the participants in the conversation take place and the implicit “rules” that they follow. The second is sequence organization: looking at the patterns and sequences of sections of the conversation that seem to relate to one another. The third is repair organization: how utterances are adjusted or corrected and occasions where the conversation has taken a different direction as a consequence of non-response. The fourth is turn construction: who speaks at any given time, who initiates aspects of the conversation, who responds, and other ways that conversational contributions may be allocated between participants.

There is not sufficient space to go into more detail here, but Ten Have (2007) provides a reasonably definitive guide to conversation analysis that outlines its procedures, conventions, and relevance. Conversation analysis is a valid form of qualitative analysis, in its own right, exploring shared meanings through observable processes but with an entirely different focus than discourse analysis.

Narrative Analysis

Narrative analysis derives from the ethnographic tradition but draws influences from a wide range of post-modern approaches. It encompasses a range of methods and is primarily concerned with exploring the accounts and stories that people use in their everyday lives. Narratives can come in a range of forms. A narrative could, for example, be someone’s individual life story, drawing from both their individual account and from other data (e.g., photograph albums or souvenirs). It could be a personal narrative of a whole period of someone’s life, given in an interview or series of interviews with a researcher. It could also be a specific story about a particular event. What really distinguishes narrative analysis from other forms of qualitative analysis, however, is that the information from a single individual is treated as a self-contained unit. Where other forms of analysis might sectionalize the data, drawing out themes for comparison from a range of sources, narrative analysis treats individual accounts as complete, exploring them as a unit for meaning and implications.

Andrews et al. (2013) identify three main approaches to narrative analysis. The first is event-centered narrative research, which draws primarily from individuals’ spoken accounts of their experiences of a single event or set of events. Data for this approach are usually interview data in one form or another, so it is closely linked with the second approach: experience-centered narrative research. The difference is that the latter focuses more on people’s general experience and how they make sense of it. Data for this type of research can vary from brief interview segments to lengthy accounts of personal life histories and may include anecdote and third-hand information from the person concerned as well as their direct personal experience. Squire (2013) discusses how one of the central concepts in this type of narrative analysis is the idea of personal agency, as the individual shapes various aspects of their experience into a coherent whole.

Event-centered and experience-centered research both work on the assumption that the narrative reflects or represents internal interpretations of memories, thoughts, feelings, and other phenomena. The third approach, however, emphasizes the social construction of reality and the way that people come to consensual narratives through conversation and other forms of social interaction. The purpose of co-constructed narrative analysis of this type is to explore the social functions of stories, though their patterns and emphases. Data for this type of analysis may involve segments or recordings of conversations, media accounts, and interviews or collections of cultural anecdotes (e.g., Georgakopoulou, 2007).

The process of conducting a narrative analysis involves the identification of themes running through the data, much as it does with other forms of qualitative analysis. These themes are used to explore the development of narrative within the single case, rather than for comparison between cases. The idea is to develop a predictive explanation of the stories in the data, cycling reflexively and repeatedly from specific example to generalizations and back again. The aim is not to develop a single "truth" but to build a case for a particular interpretation that may then be evaluated by others or compared with different interpretations. For those interested in discovering more about narrative research, Andrews et al. (2013) provide an illustrative account of the processes involved with examples from many researchers in the field.

Framework Analysis

Framework analysis is an approach to dealing with qualitative data that was developed primarily for social policy research. As such, its main aim is to provide guidance for social policy or in similar applied contexts rather than to generate a reliable or comprehensive theoretical account. Because of this focus, it has become a popular method in contexts such as consumer research or strategic management, sometimes used alone but more often contributing to a mixed-methods approach.

The data used in framework research usually come from interviews or focus groups but sometimes also from participant observations (Cresswell, 2003); it may take the form of text, audio, or video records. Once these data have been obtained, the analysis proceeds, initially with three familiar stages: familiarization, identifying themes, and indexing (Ritchie & Spencer, 1994). However, there are significant differences between framework analysis and more academic approaches to qualitative analysis. In framework analysis, it is not considered necessary for all data to be included. Given the applied nature of the research, and the richness of qualitative information, there may be time or resource constraints that make it impractical. The selection of key data is, therefore, important, and the researcher needs to ensure that material from diverse sources, time periods, and cases is fully represented in that selection (Srivastava & Thomson, 2009).

The fourth stage in framework research involves charting the data – arranging them into charts representing the themes and the parts of the data relevant to them. This involves generating a matrix in which every participant is allocated a row; the headings and subheadings for the columns are the themes and subthemes. This process has several benefits, including how it allows the data to become better organized and structured and enables new themes to be added if they seem to be relevant. Typically, the process will enable the researcher to begin to identify how general, overarching themes are shaping the material. Another positive aspect of this approach is its transparency – it is possible to see exactly how the raw data have been interpreted (Ritchie et al., 2013).

The final stage is the mapping and interpretation of the information that has been obtained through the analysis. This is guided by six key objectives typical of qualitative analysis: defining concepts, mapping the range and nature of phenomena, creating typologies, finding associations, providing explanations, and developing strategies (Ritchie & Spencer, 2002). Using these concepts, the associations between the data themes, a priori categories, and overarching themes are represented graphically, showing how they connect and interlink. The result is a visual illustration of the central concepts and relationships that have emerged from the data. For those interested in finding out more about this method, Kiernan and Hill (2018) provide a useful example, which discusses the approach and illustrates how it was used in a study of military recruit interviews.

Vignette Analysis

A vignette is a relatively short description – usually about 200 words in length – that identifies the key issues in a particular case. It summarizes one person's perceptions of significant events, including important aspects of their contexts. As such, they can be invaluable in understanding complex human interactions. Vignette analysis has been adopted as a useful method in many types of applied research, particularly in nursing, and lends itself to both qualitative and quantitative approaches.

Vignettes are, almost by definition, subjective. However, the analysis of several vignettes allows researchers to combine interpretations in a meaningful way, permitting a more holistic approach to the study of social experience. For example, Miller et al. (1997) showed how a vignette analysis was able to clarify different experiences relating to drug addiction by drawing on accounts from the family members of addicts. The various perspectives of individuals allow the researcher to triangulate on central concepts or issues, and the commonality or otherwise of what is included and what is left out is also informative.

Sometimes, vignettes are constructed directly by the participants concerned, as in studies of professional experience in applied settings (Miles, 1990). Alternatively, they may be derived from interview data – at least two people, often three, with one of them usually being the original interviewer, will draw up a vignette for each interview. Comparison of these accounts can create insights that would be entirely missed if only one individual's interpretation was adopted.

A qualitative vignette analysis focuses on identifying recurrent themes in the vignettes; these might either be content-based, such as those concerning the administrative aspects of professional practice, or to do with the emotional valence of issues (e.g., sources of emotional support or professional anxieties). Themes can also be identified through the metaphors used to describe events or situations; a description of the working environment as a "battleground", for instance, reveals a great deal about the experience of working there.

Different vignettists often adopt different levels of description. Miller et al. (1997) identified four different styles representing different levels of abstraction that can be found in vignettes. The first is a descriptive style – factual description without inference. The second is a deductive style in which the vignettist draws inferences from the material, such as inferring underlying motives. The third is a thematic style, in which the vignettist identifies consistent themes or recurrent concerns in the material. The fourth is a speculative style, such as hypothesizing about unconscious needs or motives.

Vignettes can lend themselves to either qualitative or quantitative analysis and are useful in summarizing information from fairly large samples. Bieneck (2009) gives a useful discussion about the use of this method in psycho-legal contexts that has relevance for many other areas of social research.

Conclusion

As this chapter has shown, there are many approaches to qualitative research, applicable in different contexts and for different purposes. Although it often appears superficially attractive, qualitative analysis is rarely an easy option for students or for serious researchers. Rigorous attention to detail and technique is necessary if it is to be used effectively in data analysis. This chapter provides an overview of the major methods, but for reasons of space cannot cover the full range or deal with each in the detail required for their application.

As awareness of qualitative approaches has developed, their use has become increasingly common. The increasing popularity of qualitative analysis, both in its own right and as a supplement to quantitative analysis, means that it has become relatively easy for a researcher in the social and behavioral sciences to access relevant guidance; this chapter has provided appropriate sources for each of the methods covered. Caution is necessary, however. There are many computer packages that can aid aspects of qualitative analyses, such as theme extraction or grouping, but they can only support and not replace the need for verification and understanding from the individual researcher.

These caveats aside, qualitative analysis can provide a depth of understanding and the emergence of new perspectives that can enrich social and behavioral research. As such, it is a welcome addition to the researcher's analytical "tool kit."

References

Ahmadvand, M. (2011). Critical discourse analysis an introduction to major approaches. *Dinamika Bahasa Dan Budaya*, *5*(1), 82–90 https://doi.org/10.1007/978-3-319-12616-6_4

Andrews, M., Squire, S., & Tamboukou, M. (2013). *Doing Narrative Research*. SAGE Publications.

Atawneh, A. M. (2008). The discourse of war in the Middle East: Analysis of media reporting. *Journal of Pragmatics*, *41*(2),263–278 https://doi.org/10.1016/j.pragma.2008.05.013

Babbie, E. (1992). *The Practice of Social Research*. Macmillan.

Bartlett, F. C. (1932). *Remembering*. Cambridge University Press.

Beloff, H. (1997). Making and un-making identities: A psychologist looks at art-work. In N. Hayes (ed.), *Doing Qualitative Analysis in Psychology* (pp. 55–68). Psychology Press.

Bieneck, S. (2009). How adequate is the vignette technique as a research tool for psycho-legal research? In M. E. Oswald, S. Bieneck, & J. Hupfeld-Heinemann (eds.), *Social Psychology of Punishment of Crime*. John Wiley & Sons.

Boyatzis, R. E. (1998). *Transforming Qualitative Information: Thematic Analysis and Code Development*. SAGE Publications.

Braun, V. & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3, 77-101. https://doi.org/10.1191/1478088706qp063oa

Brown, G. & Harris, T. (1978). *The Social Origins of Depression: A Study of Psychiatric Disorder in Women*. Routledge.

Bryant, A. (2019). *The Varieties of Grounded Theory*. SAGE Publications.

Bryant, A. & Charmaz K. (2010). *The SAGE Handbook of Grounded Theory*. SAGE Publications.

Charmaz K. (2014). *Constructing Grounded Theory*, 2nd ed. Sage Publications.

Cho, J. Y. & Lee, E-H. (2014). Reducing confusion about grounded theory and qualitative content analysis: Similarities and differences. *The Qualitative Report*, *19*, 1–20. https://doi.org/10.46743/2160-3715/2014.1028

Creswell, J. W. (2003). *Research Design: Qualitative, Quantitative and Mixed Methods Approaches*. SAGE Publications.

Daghigh, A. J & Rahim, H. A. (2020). Representation of Muslim minorities in politicians' discourse: Jacinda Ardern vs. Donald Trump. *Journal of Muslim Minority Affairs*, *40*(2), 179–195. https://doi.org/10.1080/13602004.2020.1773099

Downe-Wamboldt, B. (1992). Content analysis: Method, applications, and issues. *Health Care for Women International*, *13*, 313–321. https://doi.org/10.1080/07399339209516006

Edwards, D. (1997). *Discourse and Cognition*. SAGE Publications.

Fairclough, N. (1995). *Critical Discourse Analysis*. Longman.

Georgakopoulou, A. (2007). *Small Stories, Interaction and Identities*. John Benjamins.

Gill, R. (1996). Discourse analysis: Practical implementation. In J. T. E. Richardson (ed.), *Handbook of Qualitative Research Methods*. BPS Books.

Glaser, B. G. (1998). *Doing Grounded Theory: Issues and Discussions*. Sociology Press.

Glaser B. G. & Strauss A. L. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine.

Graneheim, U. H. & Lundman, B. (2004). Qualitative content analysis in nursing research: Concepts, procedures and measures to achieve trustworthiness. *Nurse Education Today*, *24*(2),105–112. https://doi.org/10.1016/j.nedt.2003.10.001

Grounded Theory Institute (2008). What is grounded theory? Available at: www.groundedtheory.com/what-is-gt.aspx.

Hayes, N. (2021). *Doing Psychological Research*, 2nd ed. Open University Press.

Hayes, N. (ed.) (1997a). *Doing Qualitative Analysis in Psychology*. Psychology Press.

Hayes, N. (1997b). Theory-led thematic analysis. In N. Hayes (ed.), *Doing Qualitative Analysis in Psychology* (pp. 93–114). Psychology Press.

Hesse-Biber S. N. & Leavy, P. (2011). *The Practice of Qualitative Research*, 2nd ed. SAGE Publications.

Hseih, H-F. & Shannon, S. E. (2005). Three approaches to qualitative content analysis *Qualitative Health Research*, *15*(9), 1277–1288. https://doi.org/10.1177/1049732305276687

Jahoda, M., Lazarsfeld, P. F., & Zeisel, H. (1932). *Marienthal: The Sociography of an Unemployed Community*. Aldine. (English ed. 1971, Routledge).

Kiernan, M. D. and Hill, M. (2018). Framework analysis: a whole paradigm approach. *Qualitative Research Journal*, *18*(3), 248–261. https://doi.org/10.1108/QRJ-D-17-00008

Krauss, S. E. (2005). Research paradigms and meaning making: A primer. *The Qualitative Report*. *10*(4), 758–770. https://doi.org/10.46743/2160-3715/2005.1831

Masroor F., Khan, Q. N., Aib, I., & Ali Z. (2019). Polarization and ideological weaving in Twitter discourse of politicians. *Social Media and Society*. October–December, 1–14. https://doi.org/10.1177/2056305119891220.

Mead, M. (1928). *Coming of Age in Samoa*. William Morrow.

Miles, M. B. (1990). New methods for qualitative data collection: Vignettes and pre-structured cases. *International Journal of Qualitative Studies in Education*, *3* (1), 37–51. https://doi.org/10.1080/0951839900030104

Miller, T., Velleman, R., Rigby, K., et al. (1997). The use of vignettes in the analysis of interview data: Relatives of people with drug problems. In N. Hayes (ed.), *Doing Qualitative Analysis in Psychology* (pp. 201–226). Psychology Press.

Nortio, E., Varjonen, S. Mähönen, T. A., & Jasinskaja-Lahti, I. (2016). Interpretive repertoires of multiculturalism: Supporting and challenging hierarchical intergroup relations. *Journal of Social and Political Psychology*, *4*(2), 2195–3325. https://doi.org/10.5964/jspp.v4i2.639

Parker, I. (1994). Discourse analysis. In P. Banister, E. Burman, I. Parker, M. Taylor, & C. Tindall (eds.), *Qualitative Methods in Psychology: A Research Guide*. Open University Press.

Patton, M. Q. (2002). *Qualitative Research and Evaluation Methods*. SAGE Publications.

Pidgeon, N. & Henwood, K. (1997). Using grounded theory in psychological research. In N. Hayes (ed.), *Doing Qualitative Analysis in Psychology* (pp. 245–274). Psychology Press.

Pidgeon, N. & Henwood, K. (2004). Grounded theory. In M. A. Hardy & A. Bryman (eds.), *Handbook of Data Analysis*. SAGE Publications.

Potter, J. (1996). Discourse analysis and constructionist approaches: Theoretical background. In: J. T. E. Richardson (ed.), *Handbook of Qualitative Research Methods*. BPS Books.

Potter, W. J. & Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research*, *27*, 258–284. https://doi.org/10.1080/00909889909365539

Ritchie, J. & Spencer, L. (1994). Qualitative data analysis for applied policy research. In A. Bryman and R. G. Burgess (eds.), *Analyzing Qualitative Data*. Routledge.

Ritchie, J. and Spencer, L. (2002). Qualitative data analysis for applied policy research. In M. A. Huberman & M. B. Miles (eds.), *The Qualitative Research Companion*. SAGE Publications.

Ritchie, J., Lewis, J., McNaughton-Nicholls, C. & Ormston, R. (2013). *Qualitative Research Practice: A Guide for Social Science Students and Researchers*. SAGE Publications.

Ryan, G. W. & Bernard, H. R. (2000). Data management and analysis methods. In N. K. Denzin & Y. S. Lincoln (eds.), *Handbook of Qualitative Research*, 2nd ed. (pp. 769–802. SAGE Publications.

Ryan G. W. & Bernard H. R. (2003). Techniques to identify themes. *Field Methods*, *15*(1), 85–109. https://doi.org/10.1177/1525822X02239569

Sherrard, C. (1997). Repertoires in discourse: Social identification and aesthetic taste. In N. Hayes (ed.), *Doing Qualitative Analysis in Psychology* (pp. 69–92). Psychology Press.

Singh, S. & Estefan, A. (2018). Selecting a grounded theory approach for nursing research. *Global Qualitative Nursing Research*, *5*(2). https://doi.org/333393618799571

Smith, T. (2020). *The Root and Uses of Marketing Knowledge*. deGruyter.

Squire, C. (2013). From experience-centred to socially-oriented approaches to narrative. In M. Andrews, S. Squire, & M. Tamboukou (eds.), *Doing Narrative Research*. SAGE Publications.

Srivastava, A. & Thomson, S. B. (2009). Framework analysis: A qualitative methodology for applied policy research. *Journal of Administration and Governance*, *4*(2), 72–79.

Stenner, P. (1993). Discoursing jealousy. In E. Burnam & A. Parker (eds.), *Discourse Analytic Research: Repertoires and Readings of Texts in Action*. Routledge.

Strauss, A. & Corbin, J. (1990). *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*. SAGE Publications.

Suddaby, R. (2006). What grounded theory is not. *Academy of Management Journal*, *49*, 633–642. https://doi.org/10.5465/amj.2006.22083020

Ten Have, P. (2007). *Doing Conversation Analysis: A Practical Guide*, 2nd ed. SAGE Publications.

Thornberg, R. (2012). Informed grounded theory. *Scandinavian Journal of Educational Research*, *56*(3), 243–259. https://doi.org/10.1080/00313831.2011.581686

Thornberg, R. Perhamus, L. M., & Charmaz, K. (2015). Grounded theory. In O. Saracho (ed.), *Handbook of Research Methods in Early Childhood Education. Volume 1* (pp. 405–439). Information Age Publishing.

Urquhart, C., Lehmann, H., & Myers, M.D. (2010). Putting the 'theory' back into grounded theory: Guidelines for grounded theory studies in information systems. *Information Systems Journal*, *20*, 357–381. https://doi.org/10.1111/j.1365-2575.2009.00328.x

Ussher, J. M. & Mooney-Somers, J. (2000). Negotiating desire and sexual subjectivity: Narratives of young Lesbian Avengers. *Sexualities*. *3*, 183–200. https://doi.org/10.1177/136346000003002005

Vaismoradi, M., Jones, J., Turunen, H., & Snelgrove, S. (2016). Theme development in qualitative content analysis and thematic analysis. *Journal of Nursing Education and Practice*, *6*(5), 100–110. https://doi.org/10.5430/jnep.v6n5p100

Van Dijk, T. (2006). Discourse and manipulation. *Discourse and Society*, *17*, 359–383. https://doi.org/10.1177/0957926506060250

Weber, R. P. (1990) *Basic Content Analysis*. SAGE Publications. https://dx.doi.org/10.4135/9781412983488

Wodak, R. (2001). What CDA is about: A summary of its history, important concepts, and its development. In R. Wodak and M. Meyers (eds.), *Methods of Critical Discourse Analysis*. SAGE Publications. https://dx.doi.org/10.1057/9780230288423_3

PART V

Tips for a Successful Research Career

29 Designing a Line of Research

Sheldon Solomon, Jeff Greenberg, and Tom Pyszczynski

Abstract

Finding one’s niche in any scientific domain is often challenging, but there are certain tips and steps that can foster a productive research program. In this chapter, we use terror management theory (TMT) as an exemplar of what designing a successful line of research entails. To this end, we present an overview of the development and execution of our research program, including testing of original hypotheses, direct and conceptual replications, identification of moderating and mediating variables, and how efforts to understand failures to replicate mortality salience effects led to important conceptual refinements of the theory. Our hope is that recounting the history of terror management theory and research will be useful for younger scholars in their own research pursuits in the social and behavioral sciences.

Keywords: Programmatic Research, Theory Development, Hypothesis Testing, Replication, Theoretical Refinement

Introduction

In this chapter, we describe how we developed our research program as a means to illustrate how to develop a theory, generate testable hypotheses, and execute a systematic research program to assess the validity of the original theory and provide empirical bases for theoretical refinements. Although all research programs likely vary as a function of the nature of the questions addressed, the personal and professional predilections of the researchers, and the epistemological tenor of the times, our hope is that recounting the history of terror management theory and research will yield some general insights and strategies useful for those currently embarking on, or enmeshed in, their own theoretical and empirical pursuits.

Terror management theory (TMT; Greenberg et al., 1986; Solomon et al., 1991a) was formulated in the 1980s to elucidate the psychological functions of self-esteem and cultural worldviews. Our research program, testing hypotheses derived from TMT, spans more than three decades. It has produced a corpus of evidence consistent with the core tenets of the theory, led to theoretical refinements of the theory (Pyszczynski et al., 1999), inspired new conceptualizations and research programs related to the theory – for example, Goldenberg and Arndt’s (2008) terror management health model and Pyszczynski and Kesebir’s (2011) anxiety-buffer disruption theory, and has withstood theoretical and empirical challenges (Pyszczynski et al., 2015). Moreover, TMT contributed to the development of experimental existential psychology (Greenberg et al., 2004) as a subdiscipline of psychological science.

Background and Training

We met as graduate students in the experimental social psychology PhD program at the University of Kansas (KU) in the late 1970s; as we became friends, we recognized our mutual interest in two central questions that we felt were not being adequately addressed in social psychological discourse at the time. First, what is self-esteem and why do people crave it so fervently? Although William James, in *The Principles of Psychology* (James,1890), identified self-esteem as a fundamental human need, just as essential for survival and effective functioning as biological needs for nutrition, social psychologists had not addressed the question of *why* people need self-esteem. The second question was, what is the psychological basis of prejudice and ethnic strife? Why is it so difficult to peacefully coexist with others who are different from ourselves? Why is human history a continuing succession of genocidal atrocities punctuated by the brutal subjugation of domestic inferiors? In the aftermath of Nazism and World War II, social psychology focused on this second set of questions in the hope of promoting a more peaceful and equitable world.

Our approach to these questions, and research in general, was substantially influenced by the strong emphasis on theory development that was central to our graduate training at KU. According to Kurt Lewin (1951, p. 169), “there’s nothing so practical as good theory.” Fritz Heider, one of Lewin’s most distinguished students, creator of balance theory, and author of *The Psychology of Interpersonal Relations* (Heider, 1958), was an emeritus professor at KU in our early days. Jack Brehm, who conducted the first cognitive dissonance experiment and later created reactance theory and the theory of motivational suppression, became our mentor. Jack was an early student of Leon Festinger, another famous Lewin student and prominent social psychologist, who created cognitive dissonance theory. Finding mentors who inspired creative and critical thinking was essential to our development and is something worth pursuing at all stages of one’s scholarly pursuits.

There was, moreover, a predilection toward motivational accounts of human attitudes and behavior. Though emphasis on developing motivational theories at KU was not surprising, given the history and composition of the department, our graduate education occurred at a time when social psychological discourse was moving in a more cognitive direction and a preference for multiple “mini-theories” – focused on increasingly detailed explanations for previous research findings, rather than broad psychological theories focused on pressing individual and social problems – dominated the field.

For example, Greenwald et al. (1986) argued that researchers get too enamored with their theories and, hence, ignore data inconsistent with them; moreover, theory-driven researchers are prone to tinkering with experimental procedures to render their empirical findings consistent with theoretical predictions (see Chapter 1 in this volume). They proposed a “result-centered” approach to research based on efforts to determine specific conditions under which a particular known finding can or cannot be obtained while delineating conditions under which a previously unobtainable result can be produced. We responded that Greenwald et al.'s result-centered approach would not eliminate confirmation biases, although it would "encourage increasingly narrow and trivial research endeavors and discourage the development of more useful methods and theories for understanding human behavior" (Greenberg et al., 1988, p. 566). This was consistent with Albert Einstein's view of science (Schilpp, 1979, p. 47): "Even scholars of audacious spirit and fine instinct can be obstructed in the interpretation of facts by philosophical prejudices. The prejudice ... consists in the faith that facts by themselves can and should yield scientific knowledge without free conceptual construction."

Additionally, we learned that, once a theoretical idea was developed, one should derive testable hypotheses and conduct programmatic research to assess the validity of those hypotheses and ultimately the theory from which they were derived. Rather than conducting single studies to make a proverbial splash, or initially trying to explore all the complexities of the process of interest with a complex research design, one should generate a simple hypothesis, and test it directly. If that hypothesis gains initial support, the robustness and replicability of the effect should be assessed by subsequently ruling out alternative explanations for the finding and identifying moderating conditions, with a goal of delineating mediational processes that underlie the effects in question. Additionally, one should enhance confidence in a theory's validity by conceptual replications of empirical studies employing multiple operations of independent variables and dependent variables. Conceptual replication is essential for demonstrating that findings reflect the conceptual variable of interest, rather than the particular operationalizations of those variables in any given study.

Finally, developing terror management theory and our research program was undoubtedly facilitated by the different interests and skills that we each brought to the theoretical and empirical table. Tom arrived at KU with a strong background in behaviorism and attribution theory, Jeff was well versed in the stereotype and prejudice literature, and Sheldon had a strong background in motivational theories (e.g., cognitive dissonance theory). Tom and Jeff had strong statistical and methodological skills, and Sheldon had experience collecting autonomic measures of physiological arousal and was widely read in the natural and social sciences. Assembling a research team with diverse interests and skills strikes us as increasingly important in the twenty-first century, as it becomes abundantly clear that processes of interest to the social and behavioral sciences operate at multiple levels of abstraction (see Chapter 32 in this volume).

Terror Management Theory

TMT was derived from cultural anthropologist Ernest Becker's (1971, 1973) interdisciplinary effort to integrate and synthesize theories and findings from evolutionary biology, anthropology, existential philosophy, psychology, sociology, theology, humanities, and popular culture to address the motivational question, “What makes people act the way they do?” (Becker, 1971, p. vii). The theory starts with the Darwinian assumption that humans share, with all other life forms, an evolved suite of biological predispositions that facilitate survival, ultimately in the service of gene perpetuation. However, humankind is unique in its capacity for abstract symbolic thought, which fosters the development of explicit self-awareness and facilities that the existential philosopher Soren Kierkegaard argued engender awe and dread. It is awesome to be alive and to know it, but it is also dreadful to be alive and realize that one’s death is inevitable, can occur at any moment without forewarning or prospect of avoidance, and that one is a defecating and fornicating animal no more consequential or persistent than a turtle or a turnip.

Ongoing and explicit awareness of the unvarnished reality of the human condition could result in potentially debilitating waves of existential terror that is both intensely aversive and could potentially undermine effective instrumental behavior. Following Becker, TMT posits that humans manage the potential for existential terror (hence the term “terror management”) by embracing *cultural worldviews* – humanly constructed beliefs about reality, shared by individuals in a group, which minimize death anxiety by affording a sense that one is a person of value inhabiting a world of meaning. All cultural worldviews infuse existence with meaning, order, and stability by providing an account of the origin of the universe, prescriptions for appropriate conduct, and promises of literal or symbolic immortality to those who meet or exceed those standards. *Self-esteem* – the belief that one is a significant contributor to a meaningful universe – results from fulfilling expectations associated with one’s social role in the context of one’s cultural worldview.

Finally, given that the putative function of cultural worldviews and self-esteem is to mitigate existential terror, TMT posits that people are highly motivated to maintain faith in their own cultural worldview and confidence in their self-worth from the perspective of that worldview. Consequently, threats to the integrity of either component of the dual-component anxiety buffer (cultural worldviews and self-esteem) gives rise to a potential for anxiety that leads to compensatory defensive reactions that bolster one’s worldview and fortify self-esteem.

Empirical Assessments of Terror Management Theory

Most accounts of research programs in journals and (especially) textbooks portray scientific inquiry as an orderly linear progression – start with a question, formulate a theory, derive and assess basic hypotheses to generate robust and replicable effects, rule out alternative explanations, identify theoretically relevant moderators, and delineate underlying mediational mechanisms for the effects in question. TMT research did indeed originate with this model in mind, and has over time been successful in this regard, but not without some detours and bumps in the road along the way.

We found Becker’s ideas revelatory and profound because they provided a unifying conceptual framework for answering our two seemingly disparate questions about self-esteem and prejudice. Cultural worldviews provide bases for self-esteem, and self-esteem – as a sense of enduring significance – serves to buffer anxiety in general, and thoughts about death in particular. Because cultural worldviews cannot generally be verified by direct observation, they require consensual validation for viability; consequently, encountering others with different beliefs challenges this consensus and, thereby, undermines the basis for one's own psychological equanimity. Therefore, those who espouse an alternative conception of reality are threatening and must be neutralized. Denigrating, demonizing, dehumanizing, and even destroying the "other" are typical reactions to defend and bolster faith in one's cultural worldview and self-esteem.

Most research-focused psychologists, as well as empirically oriented scholars in other social and behavioral sciences, did not share our enthusiasm for these ideas. Interestingly, we generally had warmer receptions from clinicians and other practitioners, academics in the humanities, and lay audiences. When we first introduced TMT, at the 1984 meeting of the Society for Experimental Social Psychology, the audience started drifting away as soon as we mentioned that our theory was influenced by sociology, anthropology, existential philosophy, and (especially) psychoanalysis. Around the same time, we submitted a theoretical paper to the *American Psychologist*, which was summarily rejected. One review was a single line – "I have no doubt that this paper would be of no interest to any psychologist, living or dead." We found these reviews inadequate justification for rejection, and after quibbling with various editors, Leonard Eron sent us an encouraging note along the lines of, "Your theory may have merit but will not gain valid currency until you collect empirical evidence to support it."

We then realized that our graduate school training gave us precisely the tools needed to derive hypotheses from this theory and to operationalize key variables in ways to test them. We consequently initiated a research program to assess the merits of a set of converging hypotheses derived from TMT. Meanwhile, our original paper was subsequently rejected by several journals, even after two *Journal of Personality and Social Psychology* papers presenting empirical evidence for TMT were in print. Eventually, Mark Zanna, as editor of *Advances in Experimental Social Psychology*, overruled reviewers, and a comprehensive presentation of TMT was published in 1991.

The Anxiety-Buffering Properties of Self-Esteem

Our first idea was to directly test the notion that self-esteem buffers anxiety. There was already a substantial literature consistent with the self-esteem as anxiety buffer hypothesis. First, self-esteem is positively correlated with mental and physical well-being and good performance under stress, and negatively correlated with anxieties and depression (see Solomon et al., 1991b, for a review). Second, experimentally manipulated threats to self-esteem increase anxiety and motivate defenses against these threats; employing those defenses reduces anxiety. Indeed, we had done some of these studies ourselves (e.g., Greenberg et al., 1982), but we needed to test a novel hypothesis derived from the anxiety-buffer idea; what we came up with was if self-esteem buffers anxiety, then boosting self-esteem should reduce anxiety under conditions of threat, even when the source of that anxiety is unrelated to the content of the self-esteem boost.

Accordingly, Greenberg et al. (1992b) randomly assigned participants to receive positive or neutral personality feedback and then had them view graphic depictions of death from the documentary film *Faces of Death* (1978), which included an autopsy and electrocution of a death row inmate; control participants viewed neutral nature images from the same film. Although neutral self-esteem participants showed a significant increase in self-reported anxiety in response to the death-related video (supporting the effectiveness of the threat manipulation), those who received a self-esteem boost showed no increase in self-reported anxiety in response to threat. A self-esteem scale administered at the end of the study confirmed the effectiveness of the self-esteem manipulation.

A second study replicated this finding with a different manipulation of self-esteem (a physical threat) and a physiological measure of anxiety (galvanic skin response). Specifically, after receiving positive feedback or no feedback on a supposed IQ test, participants watched a series of colored lights that some thought would signal oncoming painful electrical shocks; others were told they were simply visual stimuli. All participants perspired more in anticipation of electrical shocks than colored lights (establishing the effectiveness of the threat manipulation); however, this effect was diminished when self-esteem was augmented. A third study replicated this moderating effect of a self-esteem boost on autonomic reactions to threat of shock using positive vs. neutral personality feedback to manipulate self-esteem. These experiments supported the TMT proposition that self-esteem serves as an anxiety buffer across a converging set of manipulations of self-esteem and threat, measures of anxiety, and experimental designs.

Mortality Salience and Worldview Defense

It was not immediately obvious how to assess one of the central claims of TMT – that fear of death motivates adherence to, and bolstering of, one's cultural worldview. Becker (1973) viewed fear of death as an ongoing unconscious motivator that results from being an animal predisposed to survive while knowing death will inevitably thwart this fundamental biological imperative. This unconscious fear seemed unlikely to be tapped by self-report measures and is posited to be a constant rather a variable. The idea we came up with, now known as the mortality salience (MS) hypothesis, was a bit of a leap not inherent in Becker's ideas or the initial formulation of the theory. We wondered, if cultural worldviews and self-esteem protect people from their fear of death, maybe reminding people of their mortality would intensify their need to adhere to and bolster these psychological resources. Even if knowledge of the inevitability of death is a constant or given, perhaps the need for protection from the anxiety resulting from this knowledge increases when this awareness approaches consciousness.

Testing this hypothesis (six studies by Rosenblatt et al., 1989) was the result, as sometimes happens in scientific pursuits, of capitalizing on available resources. Deb Lyon, one of Jeff's students at the University of Arizona, was dating a municipal court judge in Tucson. She was interested in assessing variables that affect judges' decision making, and her boyfriend had agreed to help her with her research by giving questionnaires to his fellow judges. It struck us that judges are responsible for upholding the laws and morals of the culture. Thus, we thought that, if we randomly included a reminder of mortality in half the questionnaires and then had the judges make a typical judgment, we could determine if reminders of death would motivate increased efforts to uphold the worldview – in this case, harsher judgment of a moral transgressor. A typical case these judges presided over was setting bond for alleged prostitutes. With Deb's judge's assistance, we created a hypothetical case with materials that judges typically use to make such judgments.

Deb was also instrumental in helping us find a way to remind the judges of their mortality. She was enrolled in a course on death and dying, which included an assignment in which students were encouraged to write their reactions to two-open ended queries: "Please describe the emotions that the thought of your own death arouses in you." and "Write down as specifically as you can, what you think will happen to you physically as you die and once you are dead." We used these questions as a death reminder, given to half of the participants in our study, describing it as a projective personality assessment in a study focused on personality variables and legal decision making. Then, the judges read about the alleged prostitute and were asked to set bond – our dependent measure. Our TMT-based prediction was that judges reminded of their mortality would set higher bond for the alleged prostitute. In the control group, the judges set an average bond of about $50 – the norm in Tucson at the time. Judges reminded of their mortality, however, set an average bond of $455. Judges are rigorously trained to adjudicate the law in a rational and uniform fashion, yet a subtle reminder of death appeared to put a giant fist on the scales of justice.

We then directly replicated this finding in subsequent experiments with samples of introductory psychology students. In the initial follow-up, we introduced our first theoretically derived moderating variable. Because cultural worldviews are internalized sets of beliefs and values from the culture, and vary among individuals, we included a premeasure of attitudes toward prostitution. We posited that MS should increase bond only among student participants who viewed prostitution as something that should be illegal. As predicted, the MS effect replicated but only among students who agreed that prostitution should be illegal. This study was the first of many to establish that MS effects are rarely simple main effects; rather, they depend on people's core beliefs and values, which vary considerably across individuals, cultures, and subcultures.

The fact that human behavior is complex and depends on the interaction of many factors is an important lesson that holds across all of the social and behavioral sciences. Virtually *nothing about human behavior* occurs all the time, for everyone, or under all circumstances. Our approach to following up on our initial findings was stepwise and incremental. Specifically, we tackled one conceptual issue at a time, rather than trying to resolve all of them simultaneously in a single study. Because all research is open to multiple interpretations, follow-up studies that provide conceptual replications with different operationalizations of each major variable, and include additional conditions and measures, are essential to scientific progress.

With this in mind, we conducted conceptual replications that manipulated MS in various ways; in the first study, we used a death-anxiety scale as an MS induction instead of the open-ended questions. Subsequent MS inductions included writing a sentence about death, proximity to a cemetery or funeral home, gory accident footage, a word search puzzle imbedded with death words, and subliminal primes of the word “death.” Other studies showed that death reminders produce more positive reactions to those who uphold cherished values, suggesting it is both negative reactions to those who violate cultural values that are exaggerated and positive reactions to those who exemplify them as well.

An important question raised by these studies was whether MS effects were specific to the problem of death or emerged in response to reminders of any aversive or anxiety-provoking event. To address this question, we compared the effects of MS with reminders of aversive but not fatal experiences, such as giving a speech in front of a large audience, an upcoming examination, failure, being in extreme pain, having a limb amputated, or being socially excluded. Our studies consistently showed defensive response to MS but not to these other threats. Other studies similarly ruled out the possibility that the effect was due to heightened self-awareness or autonomic arousal produced by the MS induction.

Another important step was to see if the effects of MS on worldview defense extend to other important aspects of worldview threat. As Becker (1971) argued, people who subscribe to a different worldview other than one’s own implicitly, and sometimes explicitly, challenge the validity of one’s own basis of psychological security. The mere existence of people with different worldviews reminds us that there are other ways of viewing things and, consequently, that our own worldview may not be accurate. Therefore, MS should increase derogation of someone who subscribes to a worldview different from one’s own. Supporting this idea, Greenberg et al. (1990) found that MS led Christian students to more favorably evaluate fellow Christians and less favorably evaluate a Jewish student. This finding was conceptually replicated in studies showing, for example, that MS increased ingroup bias in a minimal group paradigm, but only when the group distinction was psychologically meaningful (Harmon-Jones et al., 1996).

Having established the effect of MS on different aspects of prejudice and intergroup conflict, we started assessing the impact of related potential moderator variables, starting with authoritarianism. We reasoned that high authoritarians have especially narrow and rigid worldviews and, therefore, predicted and found that MS was especially likely to increase their rejection of those with attitudes different from their own (Greenberg et al., 1992a). In another study, Greenberg et al. (1990) manipulated MS and had participants react to two essays about the USA – one very positive about the country and one very critical of it. MS increased American participants’ favorable reactions to an essay and its author that praised the USA and increased negative reactions to an anti-USA essay and its author.

To follow up on the finding that authoritarianism moderates MS effects, we reasoned that people who value tolerance and open-mindedness would counter the MS-induced tendency to derogate different others. Accordingly, we posited and found that liberal political ideology and making salient the value of tolerance mitigates MS-induced negative responses to those with worldviews different from one's own (Greenberg et al., 1992a). Subsequent studies provided further evidence that both dispositional differences in religious, political, and social beliefs and attitudes, and manipulations of the salience of particular cultural values, determine the specific direction MS effects take (e.g., Jonas, et al., 2008).

After establishing that self-esteem buffers anxiety and that MS increases cultural worldview defense, we combined these ideas by hypothesizing that momentarily elevated or chronically high self-esteem reduces defensive responses to MS. Specifically, we predicted and found that momentarily elevated, or moderately high dispositional self-esteem, *decreased* MS-induced worldview defense (Harmon-Jones et al., 1997). This is important because it demonstrates the connection between these two logically distinct hypotheses and provides further evidence of convergence across distinct lines of reasoning that follow from the theory. The broad goals served by this work are those we recommend one keep in mind when developing a program of research – vary the ways independent and dependent variables are operationalized, assess alternative explanations, and utilize the theory to generate and test ideas about moderating variables.

Replication Prior to the "Replication Crisis"

It's quite fashionable in the twenty-first century to admonish previous generations of psychologists for ignoring issues of replication, confirmation and publication biases, and related concerns. These critiques undoubtedly have merit (see, e.g., Edlund et al., 2022), but much of the rhetoric surrounding these issues is misleading. We conducted literal replications of virtually all of the research we published, usually adding additional measures or manipulations to clarify the nature of the effects we found; consequently, the conditions necessary for literal replication were typically embedded in designs that probed for potential moderators and other elements that would advance our understanding of those processes. We were far from the only researchers who did so; indeed, this was standard operating procedure. When failures to replicate occurred (as they inevitably do in all research programs), we took it as a challenge to determine why superficially similar procedures produced different outcomes. Such empirical detective work often leads to discoveries that expand one's understanding of the phenomena in question.

The first 11 MS studies produced effects that confirmed our predictions (Greenberg et al., 1990, 1992a; Rosenblatt et al., 1989). Thereafter, however, some researchers reported that they had difficulty replicating MS effects, and we also had some mixed results in our labs. Fortunately, as we investigated these findings (or lack thereof), we discovered some important moderators of MS effects. We believe it will be instructive to describe two of the most important examples.

Mortality Salience Effects and Rational vs. Experiential System

On one occasion, Sheldon went to Syracuse University when colleagues failed to replicate the basic MS effect, and noticed two seemingly inconsequential differences: our stimulus materials were haphazardly printed from purple ditto sheets that were un-centered and had some typos, whereas the same materials at Syracuse were centered, flawless, and looked considerably more “official” and professional; similarly, the researchers at Syracuse were nattily attired in “business-casual” style while researchers in our labs looked and acted like tie-dyed hippies. Around the same time, Jeff noticed that some of our undergraduate experimenters consistently got supportive results when running MS studies while others consistently did not. The “successful” experimenters tended to dress and act casually and delivered their memorized scripts in a relaxed manner. The experimenters who did not yield supportive results tended to dress and deliver their scripts more formally.

This led us to the idea that these differences in formality engendered fundamentally different mindsets for participants in the experiments and that MS effects emerge only when participants are in a relaxed intuitive mindset rather than a more structured analytic mode of thinking. What we had in mind was Epstein’s (1983, 1985) cognitive-experiential self-theory that distinguished between rational and experiential cognitive systems. The rational system is deliberative, effortful, abstract, and primarily linguistic – it operates actively and consciously, based on logic and evidence, primarily when situational cues suggest the need for careful analysis. The experiential system, in contrast, is the default and dominant system in most circumstances – characterized by automatic and rapid preconscious information processing that seems self-evidently valid.

We hypothesized that stiff, formal experimenters and experimental trappings were putting participants in a rational mindset, perhaps because of evaluation apprehension, making them think that there were “right” answers on the manipulations and dependent measures (see Chapter 11 in this volume). On the other hand, the informal attire and demeanor of our more “effective” experimenters relaxed participants, so they approached the study in their default experiential mindset. Perhaps, then, MS effects emerged only when people were in this experiential state of mind. This prediction was supported by Simon et al. (1997). In one study, we utilized one of the experimenters who consistently ran studies that found MS effects, a young woman who dressed very informally, sat on the desk, and used her hands expressively as she talked. In the informal condition, we simply let her be herself. However, in a condition designed to elicit a more analytic and rational mindset, she dressed more formally, sat stiffly behind a desk, and delivered instructions in a formal manner. Then, after an MS or aversive control indication, all participants evaluated a pro-USA or anti-USA author. Results revealed an effect of MS on worldview defense in the informal- but not in the formal-experimenter condition.

We then conceptually replicated this study with written instructions designed to induce either a rational or experiential mindset (Kirkpatrick & Epstein, 1992). Prior to the MS or aversive control induction, rational-mode participants were instructed to “carefully consider your answers to (the questions) before responding . . . be as rational and analytic as possible in responding to these questions." Experiential-mode participants were instructed to "respond to (the questions) with your first, natural response. We are just looking for people's gut-level reactions to these questions." As predicted, MS increased cultural worldview defense but only in the experiential mindset condition.

These studies, originally undertaken to understand why MS effects were sometimes found and other times not, added experiential mindset as an important, but previously unconsidered, moderator. Thereafter, we trained our experimenters to be natural and informal and have included the experiential mindset instructions in all subsequent MS studies; we have encouraged others to do the same. The broad point here is that, in many if not most cases, when a prior finding is not replicated, the first hypothesis to consider should not be that the original finding was spurious or the product of inadvertent or intentional confirmation bias. Rather, inconsistent findings call for a careful, theory-guided consideration of possible differences in the samples, procedures, and operationalizations of the variables in the studies that might account for the discrepancy. Indeed, many of the "many labs" failures to replicate have turned out to be the result of sometimes known and sometimes not previously known moderating variables that account for the apparent "failures to replicate" (e.g., Luttrell et al., 2017; Noah et al., 2018). When embarking on programmatic research, it is important to realize that inconsistent results are part of the territory you will be exploring.

The Cognitive Architecture of Death Denial

Another failure to replicate a MS effect resulted in the development of the dual-process model of defensive reactions to conscious and non-conscious death thoughts (Pyszczynski et al., 1999). Specifically, in 1992, German thanatologist Randolph Ochsmann informed us that he was unable to replicate the usual MS effects in Germany. When we asked him to provide details of his procedures, we quickly noticed a big difference. Our MS induction never seemed to elicit much deep thought; participants tended to respond with one or two short sentences or phrases for each of the two items. In addition, affect measures provided no evidence that it increased negative affect or distress. In contrast to our induction, Randolph had created a much more potent and elaborate death reminder in which participants went through an intensive 20-minute guided fantasy concerning their imminent death from a terminal disease.

This stark difference led us to two possible explanations for the discrepant results. One was that a deeper consideration of death does not elicit the kind of defenses we were finding – it might encourage some kind of acceptance. The other was that the terror management defenses may only occur if death thoughts are no longer in people's current focal attention. What led us to this latter idea was the observation that our MS induction was subtle and our studies always had intervening instructions and measures between the MS manipulation and dependent variables. Perhaps, participants were no longer consciously thinking about death when completing the dependent measures.

In the first of several studies designed to assess these ideas, we compared our typical subtle MS induction with a deeper MS induction – the typical induction was fortified with instructions for participants to get in touch with their emotions about dying by imagining that they were terminally ill with cancer. We found the usual worldview defense in response to our relatively subtle MS induction but not the deeper one (Greenberg et al., 1994). One potential explanation for this discrepancy was that the deeper MS induction fostered greater emotional expression or death acceptance in response to death awareness, and this had a cathartic effect that diminished the subsequent need for a defensive response; this seemed unlikely because there was no evidence that the deeper MS induction altered negative affect or that negative affect was correlated with worldview defense.

This moved us toward the second explanation – that the degree of participants' explicit awareness of death thoughts, when worldview defense was assessed, was the moderating variable. In response to our subtle MS induction, participants were probably explicitly aware of death initially, but such thoughts may have dissipated or been actively suppressed in the 3–5 minutes before the worldview-defense measure was obtained. On the other hand, the deeper MS induction may have been potent enough that death thoughts remained salient and in conscious attention when the worldview defense was measured. In other words, these findings made us suspect that worldview defense in response to MS occurs only when thoughts of death are no longer in focal attention. This would be consistent with Becker's (1973) idea that the fear of death is a powerful unconscious motivational force.

To assess this possibility, we conducted an experiment in which, after an MS induction, participants engaged in a word-search task for three minutes. Some participants were instructed to search for neutral words (e.g., *drama*, *comedy*, and *cable*); others were instructed to search for death-related words (e.g., *corpse*, *burial*, and *blood*). All participants then completed the assessment of worldview defense. Results indicated that MS increased worldview defense only for participants who searched for neutral words in the three-minute period between the MS induction and worldview-defense measure. Those who searched for death-related words did not show this effect, presumably because explicit death thoughts were still on their minds when worldview defense was assessed.

Another study replicated this finding with two additional MS conditions, where the three-minute word-search task either started with neutral words and ended with death-related words or started with death-related words and ended with neutral ones. This controlled for the amount of time spent thinking of death. Worldview defense in response to MS was found in all conditions except those where death thoughts were explicit just prior to obtaining the dependent measure. This showed that the relevant factor was not how much time one spent explicitly thinking about death, but rather, whether death was in explicit awareness or not when the dependent measure was completed; worldview defense in response to MS only occurred when death thoughts were no longer in current focal awareness.

These findings suggested that the classic MS induction might instigate an immediate suppression of death-related thought that keeps the accessibility of such thoughts low. However, over time, this suppression is relaxed, resulting in a delayed increase in death-thought accessibility (DTA), which then increases worldview defense. To explore this possibility, our next study measured DTA either immediately after an MS induction or after a delay and distraction. Based on a method developed by Gilbert and Hixon (1991) to assess construct accessibility, DTA was assessed by having participants complete a set of 20-word stems – six of which could be completed as either neutral or death-related words. For example, C O F F _ _ could be completed as either *coffee* or *coffin*; other death-related words were grave, dead, skull, corpse, and stiff. Results clearly supported the suppression-based account – DTA was low immediately after an MS induction and increased after a delay and distraction (both relative to a control condition). Because this sequence of initial low DTA, which increased over time, matched the pattern from previous studies – MS not producing an increase in worldview defense until after a delay or distraction, these findings provided initial support for our speculation that MS effects occur primarily when thoughts of death are highly accessible but outside of consciousness.

The next series of studies (Arndt et al., 1997b) was aimed at directly assessing the idea that the initial response to reminders of death is to suppress such thoughts, and that worldview defense emerges only later, after this suppression is relaxed and death-related thoughts become more accessible (but remain outside of conscious awareness). In the first experiment, a typical MS induction or aversive control induction was followed by an initial DTA measure – a short distraction passage chosen because it is mundane and makes no reference to death or existential issues – and then a second DTA measure. Because a considerable body of research had shown that cognitive load undermines the effectiveness of thought suppression (Wegner, 1994), we manipulated cognitive load with a procedure developed by Gilbert and Hixon (1991) and then assessed DTA while participants were under either high or low load. Results replicated previous findings when cognitive load was low at the time DTA was assessed – DTA was *low* immediately after MS and increased over time. This is consistent with our suggestion that the initial response to MS is active suppression of death-related thoughts. However, DTA was *high* when cognitive load was high, presumably because thought suppression requires cognitive resources that were unavailable to participants in the high-cognitive-load conditions. The elevated levels of DTA under high-load conditions reflect high-cognitive-load participants' inability to suppress such thoughts after the MS induction.

A second study then assessed cultural worldview defense in response to MS, either immediately or after a delay and distraction, and under either low or high cognitive load. When cognitive load was low, results were consistent with previous findings – MS did not produce increased worldview defense immediately after MS but did produce increased worldview defense that emerged after a delay and distraction. However, when cognitive load was high – reducing participants' ability to suppress DTA – high worldview defense emerged immediately after MS. This finding suggested to us that worldview defense in response to MS is associated with (and perhaps ultimately engendered by) high levels of implicit DTA. Consistent with this notion, a third study found that DTA is reduced in the aftermath of a MS induction if participants are given an opportunity to engage in worldview defense. Then Arndt et al. (1997a) added convergent support for the notion that increased DTA, outside of conscious awareness, is a sufficient (and perhaps necessary) condition for MS-induced worldview defense; they showed that subliminal reminders of death (i.e., 28-millisecond exposure) produced both immediate increased DTA and (in another study) increased worldview defense.

These findings provided converging evidence for a dual-process model of proximal and distal defenses in response to conscious and non-conscious thoughts of death (Pyszczynski et al., 1999; depicted in Figure 29.1). Proximal defenses are instigated when thoughts of death are in focal attention, and people cope with them in a seemingly rational manner that "makes sense" and directly addresses the problem. For example, when consciously thinking about death, people might deny their vulnerability, exaggerate their health and hardiness, or simply suppress such thoughts. This enables them to convince themselves that death is a problem for the distant future and of little current relevance.

Though proximal defenses remove death thoughts from focal attention, they cannot negate the fact that death is inevitable. The fact that we will someday die is declarative knowledge that rarely garners explicit attention but is readily brought to mind. After death-related thoughts are suppressed, they rebound, lingering on the fringes of consciousness, in a state of high accessibility. This high level of implicit DTA then instigates distal defenses to bolster faith in one's cultural worldview and efforts to boost self-esteem. Distal defenses are the core components of the anxiety-buffering system specified by TMT that prevent death-related thoughts from entering conscious attention by reducing their accessibility before they can reach consciousness. In direct support of this model, Greenberg et al. (2000) demonstrated that, immediately after an MS induction, people engage in proximal defenses (vulnerability-denying defensive distortions) but showed no evidence of distal defense (exaggerated regard and disdain for similar and dissimilar others, respectively); as expected, distal defense was exhibited after a delay, but proximal defenses were not.

The sequence of inconsistent findings and conceptual puzzles, leading to the development and testing of the dual-process model of proximal and distal defenses, illustrates how unexpected findings and failures to replicate can stimulate a deeper understanding of the phenomena one is hoping to explain. To be sure, after publishing our first few papers demonstrating MS effects, learning that these findings were not being replicated by another researcher was disheartening. We wondered if our findings might have been flukes or due to errors of some sort, but we suspected there was something different about our studies that found effects of death reminders and our colleagues' studies that

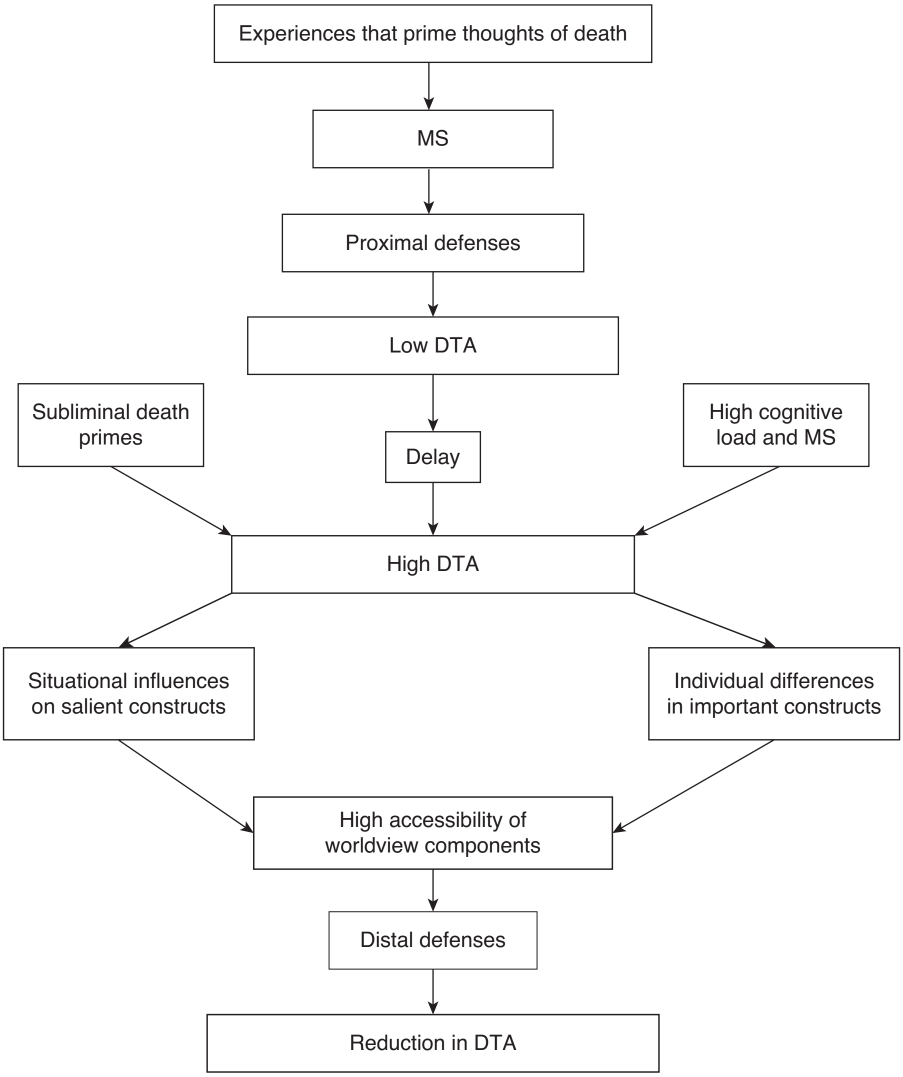

Experiences that prime thoughts of death
MS
Proximal defenses
Low DTA
Subliminal death primes
Delay
High cognitive load and MS
High DTA
Situational influences on salient constructs
Individual differences in important constructs
High accessibility of worldview components
Distal defenses
Reduction in DTA

Figure 29.1 *Proximal and distal defenses in response to conscious and unconscious death thoughts.*

did not. This led to serious discussion within and between our lab groups and an initial study that directly compared the procedures we were employing. This yielded the unexpected finding that milder MS inductions produced effects on worldview defense that stronger ones do not – a reversal of the usual dose–response relationship – and a lot of thought and discussion about why that may be the case. Eventually the pieces fell into place.

Through this process, we realized that, although we were enamored with the rather counter-intuitive idea that thoughts of death affect behavior unrelated to death – the pursuit of faith in one's cultural worldview and self-esteem – people did a lot of other things to cope with death anxiety that were more obviously related to the problem of death. They deny their vulnerability to things that could kill them, engage in health-promoting behaviors (or at least promise to do so), and often suppress or avoid thoughts of death-related issues. Research conducted to understand why what initially seemed like minor inconsequential procedural differences led to divergent findings eventually yielded a coherent conceptual picture. The general point here is that whatever it is you are studying is probably more complicated than you initially expected – inconsistent findings and failures to replicate should be taken as cues and clues to help you understand this complexity and expand your conceptualization to capture that complexity.

Transition from *Paradigm Shift* to *Normal Science*

In a sense, this was a turning point in our TMT research program. We developed TMT to explain the psychological functions of self-esteem and the psychological underpinnings of prejudice. The initial lines of TMT research were straightforward derivations of Becker's claims about the anxiety-buffering qualities of self-esteem and death-denying aspects of cultural worldviews. Thereafter, resolving discrepancies between methods and procedures used by other researchers having difficulties replicating our findings led to important theoretical refinements (e.g., that MS effects are manifested when the experiential system, but not the rational system, is engaged; see Greenberg et al., 1997 for a complete account); the theory was extended to include the dual-process model of proximal and distal defenses in response to conscious and non-conscious death thoughts.

In terms of Thomas Kuhn's (1962) depiction of science in *The Structure of Scientific Revolutions*, TMT and research reflected, at least to some extent, a *paradigm shift* in social psychology – social psychologists now accepted, or even embraced, the notion that sophisticated experimental methods could be employed to address existential questions previously deemed to be beyond the scope of empirical inquiry. Thereafter, Kuhn proposed, comes a period of *normal science*, characterized by research to establish moderating conditions that yield refinements of the original theory, using the theory to derive and test hypotheses beyond the scope of the original theoretical formulation, intersection, and possibly integration with, other relevant theories, and derivation of new areas of inquiry or research paradigms.

As support for the core tenets of TMT was emerging, all these features of the Kuhnian phase of normal science on these issues ensued. For example, Mikulincer et al. (2003) proposed, and empirically corroborated, a theoretical juxtaposition of terror management theory with attachment theory. Arndt and Goldenberg extended the dual defense model of proximal and distal defenses to health-related attitudes and behavior. The resultant terror management and health model (Goldenberg & Arndt, 2008) has generated an impressive body of empirical support with a host of important applied implications. This work suggests that, when health concerns bring death into focal attention, variables like level of optimism regarding strategies to mitigate the threat determine whether people will engage in proximal defenses that enhance or hamper their health. In contrast, when such reminders increase DTA outside of focal attention, people's worldviews and bases of self-worth determine whether their distal defenses will enhance or hinder their health. For example, immediately after an explicit death reminder, residents of South Florida, who read about the dangers of too much exposure to the sun, reported that they would use a more powerful sunscreen and spend less time at the beach. However, five minutes after an explicit death reminder, participants who base their self-esteem on their appearance reported that they would use a less powerful sunscreen and spend more time at the beach (Routledge et al., 2004).

Additionally, our attempts to understand the cognitive underpinnings of MS effects led to the generation of a new hypothesis that has yielded important evidence regarding core aspects of TMT. Specifically, the *death-thought accessibility* hypothesis states that, if cultural worldviews and self-esteem buffer potential terror engendered by the awareness of the inevitability of death, threatening a person's worldview or self-esteem should bring implicit death thoughts more readily to mind. Support for this hypothesis was provided by studies showing that DTA increased when Christian fundamentalists were confronted with logical inconsistencies in the bible (Friedman & Rholes, 2007), Canadians read an article criticizing their country (Schimel et al., 2007), and participants received negative feedback about their intelligence, were told their personality is incompatible with their career aspirations, or that they were ill prepared to give an upcoming speech (Hayes et al., 2008).

Though much has been learned about the role that death plays in life from research programs exploring different aspects of TMT, there is much more that is not yet well understood. What determines which particular aspect of a person's anxiety-buffering system is brought to bear when the problem of death comes to the forefront? How do proximal and distal defenses interact and influence each other? How do conscious beliefs about death affect the way one responds to unconscious death ideation? What is the impact of tactics for managing death anxiety beyond the scope of worldviews, self-esteem, and attachments (e.g., meditation, awe, and mystical experiences)? These are just a few questions that go beyond our existing knowledge, which will hopefully provide fertile grounds for future research programs. Science is an ongoing cumulative enterprise in which current knowledge begets new questions. You can be confident that you have designed a good line of research when others seek to answer these questions.

Recommendations for Researchers

Our hope in writing this chapter is that describing the development of our TMT research program will be useful for other scholars at various stages of their scientific endeavors. As we reflect on our experience, we are grateful for having been trained by some of the finest and most original social psychologists in our discipline. We recognize that our predilection for overarching theoretical accounts of human attitudes and behavior framed in motivational terms is an enduring remnant of our original training. Our practice of starting with simple studies to establish a finding, followed by stepwise incremental studies thereafter to replicate and extend the finding also contributed. We also appreciate that our thorough acquaintance with social psychological discourse at the time, particularly cognitive dissonance, attribution theory, self-esteem, prejudice, and stress, provided a good foundation of measures, methods, and experimental designs, which we were able to exploit for our purposes. Keeping up with classic and emerging literatures, both within and beyond one's own area of specialization, is an essential resource for fueling the development of one's own research program.

In sum then, here is our advice for designing a line of research. Pick and pursue questions that are of genuine interest to you. Be intimately acquainted with the current relevant literatures for theoretical, empirical, and methodological guidance, but do not be constrained or constricted by prior precedent or current practices. For example, researchers had been using death-anxiety scales for decades solely as a dependent measure. We used the same scales as an independent variable to make one's death momentarily salient. Start simple – don't try to do too much in a single study. Respectful engagement with constructive criticism is important and productive (we're still working on this one!). Though no one likes criticism, and sometimes critics miss the mark, critical assessment by one's peers is absolutely essential for scientific progress. Perhaps even more important and productive are efforts to resolve discrepancies when theoretically important empirical findings are apparently not replicated or are challenged by proposed alternative explanations. We hope that, just as we have been, you will be fortunate enough to receive encouragement from people you respect, meet and exchange ideas with scholars around the world (and see lots of the world too), and work with talented colleagues and students (and stellar humans who become your friends, collaborators, and scholars in their own right). If you do this, you should be able to reflect back after a long career, thinking that you learned a lot, accomplished enough to push back the frontiers of science a bit (even a just noticeable difference will suffice!), while still asking how the ideas and research could be refined and improved. Of course, don't forget to have lots of fun along the way!

References

Arndt, J., Greenberg, J., Pyszczynski, T., & Solomon, S. (1997a). Subliminal exposure to death-related stimuli increases defense of the cultural worldview. *Psychological Science*, *8*(5), 379–385. https://doi.org/10.1111/j.1467-9280.1997.tb00429.x

Arndt, J., Greenberg, J., Solomon, S., Pyszczynski, T., & Simon, L. (1997b). Suppression, accessibility of death-related thoughts, and cultural worldview defense: Exploring the psychodynamics of terror management. *Journal of Personality and Social Psychology*, *73*(1), 5–18. http://www.ncbi.nlm.nih.gov/pubmed/9216076

Becker, E. (1971). *The Birth and Death of Meaning: An Interdisciplinary Perspective on the Problem of Man*, 2nd ed. Free Press.

Becker, E. (1973). *The Denial of Death*. Free Press.

Edlund, J. E., Cuccolo, K., Irgens, M. S., Wagge, J. R., & Zlokovich, M. S. (2022). Saving science through replication studies. *Perspectives on Psychological Science*, *17*(1), 216–225. https://doi.org/10.1177/1745691620984385

Epstein, S. (1983). The unconscious, the preconscious and the self concept. In J. Suls & A. Greenwald (eds.), *Psychological Perspectives on the Self*, (vol. 2, pp. 219–247). Erlbaum.

Epstein, S. (1985). The implications of cognitive-experiential self theory for research in social psychology and personality. *Journal for the Theory of Social Behavior*, *15*, 283–310.

Friedman, M. & Rholes, W. S. (2007). Successfully challenging fundamentalist beliefs results in increased death awareness. *Journal of Experimental Social Psychology*, *43*(5), 794–801. https://doi-org.lib-proxy01.skidmore.edu/10.1016/j.jesp.2006.07.00

Gilbert, D. T. & Hixon, J. G. (1991). The trouble of thinking: Activation and application of stereotypic beliefs. *Journal of Personality and Social Psychology*, *60*(4), 509–517. https://doi-org.lib-proxy01.skidmore.edu/10.1037/0022-3514.60.4.509

Goldenberg, J. L. & Arndt, J. (2008). The implications of death for health: A terror management health model for behavioral health promotion. *Psychological Review*, *115*(4), 1032–1053. https://doi-org.lib-proxy01.skidmore.edu/10.1037/a0013326

Greenberg, J., Pyszczynski, T., & Solomon, S. (1982). The self-serving attributional bias: Beyond self-presentation. *Journal of Experimental Social Psychology*, *18*, 56–67.

Greenberg, J., Pyszczynski, T., & Solomon, S. (1986). The causes and consequences of a need for self-esteem: A terror management theory. In R. F. Baumeister (ed.), *Public Self and Private Self* (pp. 189–212). Springer-Verlag.

Greenberg, J., Solomon, S., Pyszczynski, T., & Steinberg, L. (1988). A reaction to Greenwald, Pratkanis, Leippe, and Baumgardner (1986): Under what conditions does research obstruct theory progress? *Psychological Review*, *95*(4), 566–571. https://doi-org.lib-proxy01.skidmore.edu/10.1037/0033-295X.95.4.566

Greenberg, J., Pyszczynski, T., Solomon, S., et al. (1990). Evidence for terror management theory II: The effects of mortality salience on reactions to those who threaten or bolster the cultural worldview. *Journal of Personality and Social Psychology*, *58* (2), 308–318. https://doi-org.lib-proxy01.skidmore.edu/10.1037/0022-3514.58.2.308

Greenberg, J., Simon, L., Pyszczynski, T., Solomon, S., & Chatel, D. (1992a). Terror management and tolerance: Does mortality salience always intensify negative reactions to others who threaten one's worldview? *Journal of Personality and Social Psychology*, *63*(2), 212–220. https://doi-org.lib-proxy01.skidmore.edu/10.1037/0022-3514.63.2.212

Greenberg, J., Solomon, S., Pyszczynski, T., et al. (1992b). Why do people need self-esteem? Converging evidence that self-esteem serves an anxiety-buffering function. *Journal of Personality and Social Psychology*, *63*(6), 913–922. https://doi-org.lib-proxy01.skidmore.edu/10.1037/0022-3514.63.6.913

Greenberg, J., Pyszczynski, T., Solomon, S., Simon, L., & Breus, M. (1994). Role of consciousness and accessibility of death-related thoughts in mortality salience effects. *Journal of Personality and Social Psychology*, *67*(4), 627–637.

Greenberg, J., Solomon, S., & Pyszczynski, T. (1997). Terror management theory of self-esteem and cultural worldviews: Empirical assessments and conceptual refinements. In M. P. Zanna (ed.), *Advances in Experimental Social Psychology* (vol. 29, pp. 61–139). Academic Press. https://doi-org.lib-proxy01.skidmore.edu/10.1016/S0065-2601(08)60016-7

Greenberg, J., Arndt, J., Simon, L., Pyszczynski, T., & Solomon, S. (2000). Proximal and distal defenses in response to reminders of one's mortality: Evidence of a temporal sequence. *Personality and Social Psychology Bulletin*, *26*(1), 91–99. https://doi-org.lib-proxy01.skidmore.edu/10.1177/0146167200261009

Greenberg, J., Koole, S. L., & Pyszczynski, T. (eds.) (2004). *Handbook of Experimental Existential Psychology*. Guilford Press.

Greenwald, A. G., Pratkanis, A. R., Leippe, M. R., & Baumgardner, M. H. (1986). Under what conditions does theory obstruct research progress? *Psychological Review*, *93*(2), 216–229. https://doi-org.lib-proxy01.skidmore.edu/10.1037/0033-295X.93.2.216

Harmon-Jones, E., Greenberg, J., Solomon, S., & Simon, L. (1996). The effects of mortality salience on intergroup bias between minimal groups. *European Journal of Social Psychology*, *26*(4), 677–681. https://doi.org/10.1002/(SICI)1099-0992(199607)26:4<677::AID-EJSP777>3.0.CO;2-2

Harmon-Jones, E., Simon, L., Greenberg, J., et al. (1997). Terror management theory and self-esteem: Evidence that increased self-esteem reduces mortality salience effects. *Journal of Personality and Social Psychology*, *72*(1), 24–36.

Hayes, J., Schimel, J., Faucher, E. H., & Williams, T. J. (2008). Evidence for the DTA hypothesis II: Threatening self-esteem increases death-thought accessibility. *Journal of Experimental Social Psychology*, *44*(3), 600–613. https://doi.org/10.1016/j.jesp.2008.01.004

Heider, F. (1958). *The Psychology of Interpersonal Relations*. John Wiley & Sons.

James, W. (1890). *The Principles of Psychology*. Henry Holt and Company.

Jonas, E., Martens, A., Niesta, D., et al. (2008). Focus theory of normative conduct and terror management theory: The interactive impact of mortality salience and norm salience on social judgment. *Journal of Personality and Social Psychology*, *95*, 1239–1251.

Kirkpatrick, L. A. & Epstein, S. (1992). Cognitive-experiential self-theory and subjective probability: Further evidence for two conceptual systems. *Journal of Personality and Social Psychology*, *63*(4), 534–544. https://doi-org.lib-proxy01.skidmore.edu/10.1037/0022-3514.63.4.534

Kuhn, T. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press.

Lewin, K. (1951). Problems of research in social psychology. In D. Cartwright (ed.), *Field Theory in Social Science: Selected Theoretical Papers* (pp. 155–169). Harper & Row.

Luttrell, A., Petty, R. E., & Xu, M. (2017). Replicating and fixing failed replications: The case of need for cognition and argument quality. *Journal of Experimental Social Psychology*, *69*, 178-183.

Mikulincer, M., Florian, V., & Hirschberger, G. (2003). The existential function of close relationships: Introducing death into the science of love. *Personality and Social Psychology Review*, *7*(1), 20–40. https://doi-org.lib-proxy01.skidmore.edu/10.1207/S15327957PSPR0701_2

Noah, T., Schul, Y., & Mayo, R. (2018). When both the original study and its failed replication are correct: Feeling observed eliminates the facial-feedback effect. *Journal of Personality and Social Psychology*, *114*(5), 657–664.

Pyszczynski, T. & Kesebir, P. (2011). Anxiety buffer disruption theory: A terror management account of posttraumatic stress disorder. *Anxiety, Stress & Coping: An International Journal*, *24*(1), 3–26. https://doi.org/10.1080/10615806.2010.517524

Pyszczynski, T., Greenberg, J., & Solomon, S. (1999). A dual-process model of defense against conscious and unconscious death-related thoughts: An extension of terror management theory. *Psychological Review*, *106*(4), 835–845. https://doi-org.lib-proxy01.skidmore.edu/10.1037/0033-295X.106.4.835

Pyszczynski, T., Solomon, S., & Greenberg, J. (2015). Thirty years of terror management theory: From genesis to revelation. In J. Olson & M. Zanna (eds.), *Advances in Experimental Social Psychology* (vol. 52, pp. 1–70), Elsevier.

Rosenblatt, A., Greenberg, J., Solomon, S., Pyszczynski, T., & Lyon, D. (1989). Evidence for terror management theory I: The effects of mortality salience on reactions to those who violate or uphold cultural values. *Journal of Personality and Social Psychology*, *57*(4), 681–690. https://doi-org.lib-proxy01.skidmore.edu/10.1037/0022-3514.57.4.681

Routledge, C., Arndt, J., & Goldenberg, J. L. (2004). A time to tan: Proximal and distal effects of mortality salience on sun exposure intentions. *Personality & social psychology bulletin*, 30(10), 1347–1358. https://doi.org/10.1177/0146167204264056

Schilpp, P. A. (1979). *Albert Einstein: Autobiographical Notes*. Open Court.

Schimel, J., Hayes, J., Williams, T., & Jahrig, J. (2007). Is death really the worm at the core? Converging evidence that worldview threat increases death-thought accessibility. *Journal of Personality and Social Psychology*, *92*(5), 789–803. https://doi-org.lib-proxy01.skidmore.edu/10.1037/0022-3514.92.5.789

Simon, L., Greenberg, J., Harmon-Jones, E., (1997). Terror management and cognitive-experiential self-theory: Evidence that terror management occurs in the experiential system. *Journal of Personality and Social Psychology*, *72*(5), 1132–1146. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15982118

Solomon, S., Greenberg, J., & Pyszczynski, T. (1991a). A terror management theory of social behavior: The psychological functions of self-esteem and cultural worldviews. In M. P. Zanna (ed.), *Advances in Experimental Social Psychology* (pp. 91–159). Academic Press.

Solomon, S., Greenberg, J., & Pyszczynski, T. (1991b). Terror management theory of self-esteem. In C. R. Snyder & D. R. Forsyth (Eds.), *Handbook of Social and Clinical Psychology: The Health Perspective* (pp. 21–40). Pergamon Press.

Wegner, D. M (1994). Ironic processes of mental control. *Psychological Review*, 101, 34–52. https://doi.org/10.1037/0033-295X.101.1.34

30 Successfully Publishing Research in the Social and Behavioral Sciences

Sicong Liu and Dolores Albarracin

Abstract

To survive and prosper, researchers must demonstrate a successful record of publications in journals well-regarded by their fields. This chapter discusses how to successfully publish research in journals in the social and behavioral sciences and is organized into four sections. The first section highlights important factors that are routinely involved in the process of publishing a paper in refereed journals. The second section features some factors that are not necessarily required to publish a paper but that, if present, can positively influence scientific productivity. The third section discusses some pitfalls scholars should avoid to protect their scientific career. The last section addresses general publication issues within the science community. We also recommend further resources for those interested in learning more about successfully publishing research.

Keywords: Publication, Scientist, Research, Writing, Authorship, Peer Review, Paper, Publish or Perish

Introduction

As a form of communicating scientific knowledge, publishing is at the core of every science (Riviera, 2013). It also has important consequences for the standing of universities, other institutions, and researchers within a research community (Linton et al., 2011). The ability to publish successfully is closely related to merit evaluation, reputation, tenure and promotion, job mobility, and salary (Klingner et al., 2005; Miller et al., 2011). As an consequence, researchers around the world perceive a high pressure to publish (van Dalen & Henkens, 2012), and the century-old expression, "publish or perish," has become a research topic in and of itself (see Qiu, 2010).

Publishing scientific findings is a complex process that involves an array of factors that can affect scholarly productivity. For instance, highly productive scientists in the 1950s published articles at a rate that was 50 times higher than that of their less productive counterparts (Shockley, 1957). As the size of the scientific literature has grown at an annual rate of more than 8% since the 1950s (i.e., number of publications and cited references; Bornmann & Mutz, 2015), the gap between more and less productive researchers has further expanded in the twenty-first century (Ioannidis & Klavans, 2018). In this context, this chapter discusses factors that synergistically contribute to a scholar's scientific productivity in the social and behavioral sciences. It is oriented towards a readership of relatively junior scholars, although senior scholars may find this chapter useful as well – particularly as a tool for advising graduate students or other junior researchers on scientific publishing.

Due to the limited scope of the chapter, we focus on important aspects of the publication process, beginning with the generation of research and ending with the publication process itself. First, following the general workflow in scientific publishing, we offer suggestions on research question generation, research execution, writing, journal submission, and portfolio assembly. Second, we discuss factors that enhance scientific productivity, including grant support, emergent opportunities, and a diverse research agenda. Third, we highlight some pitfalls, such as challenges arising from research ethics and predatory journals, for junior researchers who want to invest their efforts in the most rewarding and principled scientific activities available in their fields. Finally, we conclude with general thoughts about publishing and direct readers to further resources that can supplement the scope of our chapter.

Conducting Research

Research Question Generation

Identifying specific research questions can be stressful and is reported by successful scientists as a challenge that must be acknowledged and confronted (Weinberg, 2003). The main reason for this difficulty is that problem solving in research is different from other academic experiences, including graduate-school courses. Unlike course problems, that are known to be solvable, a researcher formulating a research problem is often not certain of whether solving the problem will present a contribution or if the problem is even solvable.

Establishing whether a research problem will be perceived as an important contribution involves consultation with experts as well as a time-consuming phase of reading the literature with breadth and depth (see Chapter 4 in this volume). A good piece of advice is thus to "forgive yourself for wasting time" (Weinberg, 2003) and engage in these preparatory activities even when they feel lengthy and overwhelming. With more than three million peer-reviewed articles published every year (Johnson et al., 2018), and a new journal launched every other day (Rawat & Meena, 2014), one can easily get drowned in the ocean of scientific literature and perceive no progress towards the goal of publishing research. However, identifying creative ideas does require establishing that an idea is new given what has been said in the literature (Klingner et al., 2005). In addition, having a good grasp of the literature helps to develop multiple perspectives that contribute to attempting to solve a given scientific problem, and such perspectives facilitate "strong inference."

Strong inference is an inductive inference method that can be traced back to Francis Bacon (Platt, 1964). Applying the method consists of following three sequential steps in a repetitive fashion. The first step is to generate as many alternative hypotheses as possible on a given problem. The second involves designing critical empirical procedures (e.g., an experiment) that allow a subset of the alternative hypotheses to be tested against the other hypotheses. The third step deals with carrying out the empirical procedures to obtain clean results that help to reject one or more hypotheses, after which the researcher can restart the three-step process to narrow down the remaining hypotheses. This entire process of strong inference, which has been compared to climbing a tree with forking branches (Platt, 1964), can increase the pace and quality of scientific progress. It can also enhance publications in several ways. It can reduce bias and increase the skills of researchers who would otherwise have a single hypothesis about or single approach to a scientific problem (Chamberlin, 1897). It can also improve the chances that one's research question will be of interest to others by making multiple meaningful theoretical contributions – an advantage that reviewers and editors are likely to appreciate. Finally, this approach can increase the impact of publications by clustering them under a clear and systematic research theme that makes appreciation of the contribution easier (Klingner et al., 2005).

Research Execution

Once settled on the research question and associated hypotheses, the researcher must pay attention to a collection of factors related to research execution. Overall, optimizing these factors as much as possible, given real-world constraints, should be the rule of thumb; this leads us to consider three factors: (a) research standards against which to judge the quality of research execution, (b) signal/noise mindset, and (c) obtaining results from data. With respect to research standards, a growing number of reporting guidelines can inform us of what excellence entails within a particular field. For example, some standards include CONSORT (CONsolidated Standards of Reporting Trials) for conducting clinical trials (Moher et al., 2001), PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) for systematic reviews and meta-analyses (Moher et al., 2010), STROBE (STrengthening the Reporting of OBservational studies in Epidemiology) for observational studies (von Elm et al., 2007), STARD (STAndards for Reporting Diagnostic accuracy studies) for diagnostic accuracy investigations (Bossuyt et al., 2003), and JARS (Journal Article Reporting Standards) for research in psychology (Kazak, 2018).

For scholars striving to have successful publishing experiences, checking these guidelines, frequently required in the submission guides from journals, is necessary. First, some procedures (e.g., preregistration) in the guidelines cannot be completed after certain milestones of the research process have been reached. In these cases, failure to comply with the guidelines without reasonable justification can prevent researchers from submitting to certain journals, especially high-profile ones. Second, even when a journal is willing to consider work that departs from best practices, the experts reviewing the submission are likely to flag flaws and dismiss the research, ultimately reducing the odds of the work being accepted by a refereed journal (Rozin, 2009).

The second factor related to research execution can be described as a signal/noise mindset – the intention to maximize the ratio between the phenomenon of interest and everything else (Luck, 2014). The phenomenon of interest is the signal and can be defined as the ‘‘relatively stable, recurrent, general features of the world’’ (Haig, 2005, p. 374); identifying signal is key in all disciplines. For example, in social psychology experiments, the experimenter’s influence, often taking the form of demand characteristics, must be carefully controlled when it is not part of the phenomenon of interest (McDougall, 2015, see Chapter 11 in this volume). As another example, in cognitive neuroscience experiments involving electroencephalography, the electrode recordings resulting from eyeball movements must be controlled for to ensure that one captures post-synaptic potentials – what researchers typically care about (Luck, 2014). Although the importance of having a signal/noise mindset is clear, applying the mindset to specific research operations turns out to be challenging and unclear (Liu & Tenenbaum, 2018). Perhaps because scientists know that they cannot control all sources of noise, they often do not bother to try (Rubin, 1974), a failure responsible for many struggles in publishing research. The data collected without proper experimental control and piloting are simply too noisy and, thus, unlikely to provide meaningful and coherent findings. The sooner one can develop the signal/noise mindset and start practicing it in specific projects, the more successful one will be in publishing research.

The third factor in research execution concerns obtaining results from data analysis. In general, statistical testing helps to categorize an outcome as either positive (e.g., statistically significant) or negative (e.g., not statistically significant). In addition to the possibility of obtaining exciting positive results supporting a priori hypotheses, researchers can also get positive results that seem surprising or that contradict expectations. Unexpected positive results should receive ample attention from researchers, as groundbreaking research often comes from re-interpretation of serendipitous findings (e.g., the discovery of mirror neurons; Di Pellegrino et al., 1992; Phaf, 2020). Overall, obtaining positive results increases the potential of the research being published (Dwan et al., 2008; Murtaugh, 2002) and cited (Etter & Stapleton, 2009; Leimu & Koricheva, 2005). However, the unexpected positive findings must be transparently reported in papers instead of as supporting tailor-made post hoc hypotheses – a criticized practice known as *hypothesizing after the results are known* (HARKing; Hollenbeck & Wright, 2017; Kerr, 1998). Frequently, however, the initial studies that produce unexpected findings end up being pilots for new research designed to test new a priori hypotheses about the serendipitous observation. Whatever the case, achieving a deep understanding of one’s data is essential to advance knowledge and enjoy the intellectual benefits of what is referred to as “following the data.”

Relative to positive results, dealing with negative results in publishing is more challenging. Unlike dealing with negative results that were not hypothesized, scientists still debate whether to publish negative results that were hypothesized to be positive. The debate can be summarized as balancing a file-drawer effect – an accumulation of negative scientific findings that do not see the light – with a cluttered-office effect – poor or meaningless research that overcrowds the literature and limits scientists' ability to process the meaningful information (Nelson et al., 2012). This conundrum is again resolved by resorting to strong inference, as good scientific practices are often interdependent with each other. Strong inference alleviates the problem of HARKing and publishing negative findings through the process of alternative hypothesis generation. Alternative hypothesis generation leaves positive and negative results on equal footing and frees researchers of the personal bias from having a single hypothesis.

Writing, Rewriting, and Proofreading

After the completion of the previous steps, good writing gets one's scientific work closer to getting published. As Bem puts it, "from my own experience as an editor of an APA [American Psychological Association] journal, I believe that the difference between the articles accepted and the top 15–20% of those rejected is frequently the difference between good and less good writing" (Bem, 1995, p. 176). Consistent with the anecdote, improving writing skills (e.g., through a writing course or writing support group) is a must and has been shown to improve researchers' publication rates (McGrail et al., 2006).

To become a good scientific writer is to write with clarity and accuracy (Bem et al., 1987; see Chapter 8 in this volume). First, writing should be organized in a way that guides readers through a coherent structure; that begins with the title and abstract. Novice researchers often underestimate the role played by the title and abstract, believing that these pieces can be written at the end, almost as an afterthought. However, the readers will form an opinion of the paper as early as they can. If the first material they read is not appealing, the paper will not receive much attention afterwards. If the first material they read is not clear, they will conclude that the paper has no substance or that the authors cannot write. Therefore, a scholar should treat the title and abstract as hooks, making sure that they promote the unique aspects of the paper in an enticing way from the very beginning.

Regarding the writing of the body of a paper, one school of thought is that the information flow should follow an hourglass shape (Bem et al., 1987). According to this metaphor, the introduction starts with broad general information and the following sections gradually narrow down to a more detailed discussion of the most relevant literature; after the method section, the sections broaden out again to the more general views addressed in the discussion. However, we propose a different metaphor: going from the seed to the tree. Accordingly, an article should be organized so that readers understand the point of the paper in the first two paragraphs, and the paper later develops those arguments with more detail. That is, the seed contains all the elements including (a) the problem, (b) the benefits and costs associated with the problem for readers, and (c) the proposed current solution (McEnernehy, 2021) – even though the full introduction and ultimately the whole paper is required to develop the seed. For example, in examining a particular problem in persuasion research, beginning an article with why studying persuasion is important is typically not appropriate. Although such a frame might have been right for the first empirical study ever conducted on persuasion, the opening paragraphs of a current persuasion paper must clearly detail how the new idea helps solve a specific problem that is costly for persuasion research if unsolved. Thus, a coherent first paragraph must jump straight into serving this purpose.

Second, writing should be simple and direct. Although many journal articles usually have a readership with specialized backgrounds, those writing the article should ignore this knowledge and aim to make their writing accessible to the wider public. To achieve such a goal, one should be armed with a style manual such as Strunk's *Elements of Style* (Strunk, 2007). For many journal reviewers, Strunk's recommendations for parallel sentences, avoiding temporal propositions (e.g., "since" to denote causation), and wordiness, have become laws of writing. As a result, these reviewers will deem the writing substandard if it departs from these stylistic conventions. There is also a variety of conventions that pertain to scientific writing per se. These involve using consistent labels for constructs, making sure that variables are enumerated in the same way in different places, and ensuring that tables and figures are cognitively fluent to readers. For example, one of the authors of this chapter was taught that when one has two levels of an experimental manipulation, it is best to present the higher level of the manipulated factor (e.g., the higher level of a manipulation or the experimental group as opposed to the control group) on the left of tables and figures. This allows readers to automatically compute contrasts that reveal the effect of the manipulation in a way that is not possible when the control condition appears first.

Writing is an iterative process, which begins with a first draft and concludes with the last version of a paper that is accepted for publication (see Chapter 8 in this volume). After completing the first draft, one must rewrite/edit one's own work and make sure it meets the writing style of the journal. Rewriting requires an effort towards conciseness. Following Strunk's (2007) recommendations, readers must go through each sentence or pair of sentences to see if the same meaning could be conveyed with fewer words. However, we do not adhere to the notion that good writing involves short sentences. Too many extremely short sentences, except when used for style, as in the first sentence of a paper, produce a choppy impression that can be judged as amateurish. Writing must be concise and fluent, without over-explaining or expecting the reader to go over more details than are necessary to understand the author's point. However, conciseness and choppiness are different. Rewriting also requires deleting material – often many pages of text one has spent hours crafting. However, this exercise is necessary for coherent writing; in the process of writing a paper, the author often switches directions and needs to refocus the argument. We recommend that scientific writers do not hesitate to delete – the ultimate goal is not to save text but publish ideas and findings described in a compelling and accurate way.

Rewriting can be agonizing for several reasons (Bem, 1995). First, because the author understands what (s)he meant to say in the original writing, it is challenging to identify locations with ambiguity and logic bumps. Second, editing one's own writing entails substantial compulsiveness and attention to detail. Third, rewriting often means restructuring. To overcome these difficulties, some tips are helpful.

Invite a colleague, who knows little about one's work, to read the draft as if (s)he is the reviewer at a journal. When the colleague expresses confusion at points made in the draft, the author should take notes about those places for later revision, refraining from explaining the point or acting defensively. By definition, unclear writing in the draft is whatever the colleague points out and should be revised. Also, writers should understand that academic writing is challenging for everyone (Lee & Boud, 2003) and that more productive writers do not have more time to write or fewer commitments to non-writing activities than less productive ones (Boice & Jones, 1984). Past research suggests that establishing some formal structure in writing can be beneficial (McGrail et al., 2006); this implies that regularly sharing one's writing with co-authors and joining a writing group can be worthwhile. Moreover, an efficient writing process frequently involves multiple subtasks, such as typing and checking references. Having multiple monitors, when writing on computer, or laying out a hard copy of the draft on a big table can enhance the writing process by helping the writer manage subtasks and have a broad view of the process. Finally, reading the document out loud is often beneficial, as is listening to it using automatic reading software.

Two additional notes concern references and proofreading. Reference management software, such as EndNote, Mendeley, and BibTex, has become a necessity. Given the increasingly large number of articles published every day, managing the references has become more and more challenging and error prone. The challenge is even greater when authors intend to submit work to journals following different publication style guidelines. Reference management software releases researchers from the tedious and time-consuming work of style switching. Prior to journal submission, a scholar should never forget to proofread the work. No journal reviewers like to be copy-editors (Sternberg, 2000), and most are extremely irritated when they see that the author has not bothered to submit a clean draft. Therefore, proofreading can help all authors avoid humiliating reviews pointing to every careful mistake the author has made.

Journal Submission and Peer Review

Researchers often either have target journals in mind, before beginning a research project, or consult experienced colleagues for journal suggestions prior to journal submission, However, one can also reap the benefit of modern information technology by using the Journal/Author Name Estimator (JANE; https://jane.biosemantics.org). JANE is a digital tool for journal (as well as reviewer) selection that makes recommendations by matching information from the title and/or abstract of one's manuscript to entries in PubMed. When using JANE, one must copy and paste the title or abstract of the paper as a query and review JANE's suggestions. The current JANE offers ranked suggestions of content-matching journals, authors (who can be potential good reviewers), and articles.

Finalizing the selection of journals, however, typically involves adopting some strategy. Given all the candidate journals, writers may have at least four strategies for journal selection (Sharman, 2015). One can order the candidate journals according to their impact and choose to first "go down the ladder," submitting to the best journal first. However, publishing the work with this strategy can take a long time given that high-impact journals receive many submissions and are likely to have longer and more selective review processes. Alternatively, one can submit directly to specialist journals or megajournals – two strategies that can get one's work published more quickly but will decrease the reputational value of the publication. Megajournals emerged relatively recently with an open-access and online-only form (e.g., *PLOS ONE*; Mudrak, 2021). Many of these megajournals accept papers for publication only based on the scientific soundness rather than theoretical and practical impact; although the open format can make many of their papers highly impactful. Other than megajournals, some broad open-access journals are among the most competitive outlets for scientific research, including *Science*, *Nature*, and *Proceedings of the National Academy of Sciences*.

A final strategy to choose journals is a compromise between the "slow" and "fast" strategies. Researchers may first submit to a high-impact journal, and, if rejected, they may then go directly to a megajournal. When aiming at high-impact journals, it is usually a good idea to write an inquiry letter to the journal editor and probe the editor's view (Fowler, 1993). If the editor's feedback is not encouraging, one should move on to the next journal. In addition, given the trend of open science in recent years (Munafò et al., 2017), it may also be a good idea to follow a best-practice checklist (see Aczel et al., 2020) and share one's data and materials online; high-impact journals are likely to require this, and adherence to open-science practices can be a positive factor in the peer-review process (Wicherts & Bakker, 2012).

In response to a submission, a journal's editorial decision can be any of the following: (a) accept (typically pending minor revisions); (b) revise and resubmit; (c) reject after external review, and (d) reject without external review (Klingner et al., 2005). An "accept" decision is rare and means that the submission is admired by the reviewers and editors to the degree of requiring little or no refinement prior to publication. However, it is not a situation that junior or even senior scholars should expect (Bem, 1995). More often than not, one gets a decision in the other categories. A "reject after external review" decision is typically accompanied by extensive expert feedback. A "reject without external review" decision – likely when submitting to high-profile journals – should not discourage authors from moving to the next submission option.

Receiving any "reject" decision can, of course, be disheartening because it means that the paper will no longer be considered for publication at the journal and that the authors must seek to publish it elsewhere. However, wise authors take full advantage of the feedback in the reviews to increase their chance of publishing the work in the next submission or rejoice in the fact that the paper was rejected without delay. Clearly, nobody likes to learn that one's work is inadequate, that one's writing has ambiguities, or that one's data do not show what one believes they show. However, much like a career in the arts, an academic career depends on persistence in the face of obstacles and on judgments that are subjective. Therefore, we must all remember that everybody receives rejections and that we generally like our own work better than the reviewers do. Our goal is to be able to continue on a career that requires learning, thinking, and producing knowledge, rather than expecting regular rewards or praise. Rewards and praise will come, but they often take years and are not the reason one pursues this career.

Receiving a "revise and resubmit" decision should be taken as a favorable sign towards acceptance, with an estimated 50–90% chance of eventual acceptance depending on specific factors (Warren, 2000). Therefore, authors should carefully read the reviews; this may seem overwhelming at first but will become clearer and more manageable as one works to address the concerns. Even though authors do not have to follow every single suggestion from reviewers, and can refute criticisms in the revision, they must respond to each reviewer comment and should do so with respect and gratitude. Authors may be tempted to conclude that the reviewers disagree when each makes different suggestions, but research suggests that reviewers usually give good suggestions that do not contradict each other (Fiske & Fogg, 1992). Also, authors should understand that the editor is their ally in this revision process and will often work to strengthen the paper to enhance the journal (Bem, 1995).

Assembling a Publication Portfolio

As researchers advance in their careers, an important issue is what publication portfolio to assemble. Decisions about what and how much to publish depend largely on the standards of the discipline, personal preferences, and the university/organization at which one works. Many disciplines have certain journals in which one must publish to get tenure, but it is also important to publish one's findings even if not all papers are published in top-tier journals. For example, for researchers who rely on grants, the output of a grant will be judged based on the publications it produced. Thus, the sheer level of output will be important, and a grant that produces no papers for five years may not be renewed even if one publishes a paper in *Nature* during the sixth year of the project. One common piece of advice for researchers is to produce a plate with "meat and potatoes," where the meat is the high-level papers that will improve one's reputation and the potatoes are other papers that make complementary points and yield a coherent program of research.

The Less Routine Factors

Research Grant Support

Beyond factors that are necessary for research productivity, other factors can exert substantial influence on a researcher's output. Getting grant support for research is one such factor. At the national level, an increase in the amount of government funding has been shown to result in a higher number of research publications (Payne & Siow, 2003). At the individual level, securing grant support enhances research productivity in several ways (Klesky, 2015). First, applying for a grant requires identifying a seminal research idea, formalizing the idea into specific operations, and presenting the idea through high-quality writing – all processes that can lead to new publications even if the grant is not ultimately funded. Because a grant application usually takes a few weeks or months to write, it is often an efficient, structured way of going through processes that improve one's research.

Second, obtaining a grant is highly rewarding. It not only adds excellent lines on one's curriculum vitae but also provides solid support for implementing a large research project that is otherwise not feasible. Making resource-demanding research possible, together with the perceived merits of being awarded the grant, may enhance the researcher's reputation and impact of their work. Third, the process of obtaining grants snowballs – winning a grant increases one's chance of winning another, often bigger grant. For this reason, researchers may want to start applying for grants as early in their career as possible, and many universities offer small-scale intramural grants to prepare students and faculty for future larger ones (Klesky, 2015). For instance, Duke's Institute for Brain Sciences (https://dibs.duke.edu) estimates that, for every dollar spent on its incubator awards to its faculty members, seven dollars return to Duke through external grants. Overall, although applying for a grant can mean extra work and sometimes additional stress, it is highly cost-effective from the point of view of boosting research productivity.

Emergent Publishing Opportunities

Sometimes publication opportunities are emergent. One such situation involves invitations to publish a particular paper based on the researcher's expertise. For junior researchers, such opportunities are rare and typically happen via extended invitations from senior scholars whose reputation and credibility in the field draws the initial invitations. The odds of getting an eventual acceptance are higher for invited contributions than regular ones. Still, peer review is expected, and one should treat the submission as a regular journal submission, which may not always be accepted.

Another publication opportunity involves supervising research in areas that are of interest to particular graduate students (see Chapter 36 of this volume). Generally, universities with doctoral programs have more publications than those without doctoral programs (Schweitzer, 1988); this implies both that graduate students enhance research and that faculty members interested in publishing research seek universities with doctoral programs. Good research mentoring requires thoughtful teaching and skillful communication from the scholar and can produce win–win results. Students learn from working on projects and from the experience of the mentor, and mentors benefit because teaching forces them to improve the logic of their own research questions, which often results in publications with students (Li, 2019).

A third type of emergent publication opportunity involves paying attention to new phenomena and observing new areas where one could contribute (for the role of explaining new phenomena in science, see Haig, 2005). However, the emerging phenomenon is sometimes hard to miss, and the matter becomes whether one can adapt one's research agenda to the phenomenon. The coronavirus pandemic that began in 2019 was an excellent example of how a phenomenon can stimulate new science; to understand the pandemic and its impact, research emerged regarding COVID-19 clinical features (Vetter et al., 2020) and the vaccination decision process (Jung & Albarracin, 2021).

Diversified Agenda, Research Collaboration, and Authorship

Prolific researchers sometimes have a multi-faceted research agenda (Ilie & Ispas, 2007). A research agenda can have elements such as a phenomenon/topic, a methodology, or a theoretical approach. Having a multi-faceted research agenda allows researchers to explore different scientific directions and prevents them from reaching a publishing dead end. It can also enhance the depth of one's research on a given topic – research breadth is not necessarily independent from its depth. For instance, research in cognitive psychology often entails neuroimaging methods whose application further involves understandings about electric/magnetic fields and neuronal activity, and so on.

Being open to a diverse research agenda also facilitates collaboration with other researchers with different expertises or backgrounds. Multiple disciplines, such as economics (Hudson, 1996) and sociology (Grant & Ward, 1991), have reported steady increases in collaborative publications. Such an upward trend is further hastened by technological development (e.g., teleconferencing), which facilitates collaboration from different locations or time zones (Fisher et al., 1998). However, research collaboration can also create issues related to research ethics and integrity (Ioannidis & Klavans, 2018). For example, recent years have witnessed a fair amount of research paper retractions from scientists who authored an unfeasibly high number of publications on a yearly basis (Reich, 2009; Tramer, 2013).

In addition, the issue of authorship can be complex and highly discipline-specific. In particular, high-energy and particle physics projects entail collaboration among large international teams. The subsequent publications can have an extremely long list of authors; as of 2019, the paper with the world record has 5,154 authors (Ioannidis & Klavans, 2018; Lapidow & Scudder, 2019). In the social and behavioral sciences, we suggest following the 1998 Vancouver criteria for authorship by the International Committee of Medical Journal Editors (2018). The criteria prescribe that an individual qualifies as an author by jointly meeting four standards, including (a) conceiving the work or acquiring, analyzing, or interpreting the resulting data, (b) writing or revising the scientific manuscript, (c) approving the final version of the work, and (d) agreeing to be responsible for all the published contents. Regarding authorship order, the disciplinary conventions and research contribution scenarios vary widely and this makes it difficult to provide specific guidelines. In social and behavioral sciences, for instance, the first and the last authors in a given paper having a relatively long list of authors are usually considered as the main contributors, with the first author being a junior scholar who is a leader in execution and the last author being a senior scholar who supervises and advises the project and often rewrites the manuscript multiple times.

The Pitfalls

Violating Rules

If we can compare publishing research to playing board games, all the previously discussed factors resemble the rules of the game. However, the rule book would not be complete without those about player elimination; in science, this involves ethical norms and regulations. These rules concern (a) research with human participants, (b) publication practices, and (c) citations.

To begin, researchers should study and abide by research ethics in protecting human participants (see Chapter 2 in this volume). In the social and behavioral sciences, scholars generally follow the ethical guidelines from the Declaration of Helsinki (DoH; World Medical Association, 2008). The first version of DoH was created in 1964 by the World Medical Association in response to growing concerns about unethical medical practices during and after World War II. The DoH sets a balance between the interests of humanity and individual patients within clinical trials. Although the full document contains thousands of words, its gist is to "do no harm" to the participants (Carlson et al., 2004). The DoH is the cornerstone document for many other ethical guidelines, such as the ethical code from the American Psychological Association (2017), and is used by institutional review boards (IRBs) during the review and approval of research protocols involving human participants. All research executions must conform to the IRB-approval protocol and often must explicitly report this in their published versions. Protocol violations (e.g., failure to obtain informed consent from participants) often results in a range of consequences for the research project, the research team, and the institution.

Beyond research ethics about human participants, researchers must not violate scientific integrity in publications (Masic, 2011). The pressure to "publish or perish" has led to an increasing number of dubious research practices, including fraud, salami slicing, and duplicate publication, to name a few (Rawat & Meena, 2014). Fraud refers to reporting fabricated or falsified research outcomes in publications; salami slicing involves splitting the same research into many fragments, publishing each of them as if they are unrelated; and duplication means submitting the same material to multiple journals and avoid getting caught by plagiarism software by systemically varying the titles, keywords, and co-authors. Duplication is related to the Ingelfinger rule, named after the former editor of the *New England Journal of Medicine*, who in 1969 declared that he would not consider a manuscript for publication when the manuscript was simultaneously submitted elsewhere or published in similar forms elsewhere (Neill, 2008). All scientists should avoid dubious research practices (e.g., publishing untrustworthy findings and selectively reporting statistics) at all cost and report, to regulatory parties, if they observe possible misconduct in others.

Finally, researchers should be aware of the ethics of citing past research. Inaccurate citations in the literature have received growing attention in recent years. For instance, the general surgery literature has a 35.4% error rate in citations (Awrey et al., 2011) – terrifying considering the potential practical implications.

From the perspective of scientific development, making inaccurate citations can distort the literature by unfairly favoring unsupported ideas (Smith & Banks, 2017). Therefore, young scholars could make their first contribution to science by being careful in citing others' work. A common error is to cite the evidence that one is using to derive a hypothesis as direct support for the hypothesis. Another error is to cite research one does not understand well and results in misattributing or misdescribing past scientific findings. A final malpractice involves biases in citations, such as overly citing one's own work or citing others' work based on factors not related to research quality and content. For instance, some have heard stories about a senior scholar advising a student to remove citations from another scholar and to avoid having that person as a reviewer. In most situations, citing the work of others should be based on the relevance and merits of the research.

Writing Hazards

Learning to recognize and avoid predatory journals is not trivial. Given an exponential rise in digital journals along with a decline in paper journals, a junior scholar can easily get confused when trying to distinguish quality journals from predatory ones. This difficulty increases one's risk of being seduced by predatory journals, especially given the harsh publish-or-perish reality and marketing strategies used by those journals (Sharman, 2015). To illustrate this pitfall, one researcher cooked up a spurious article using www.randomtextgenerator.com and submitted the article to 37 open-access journals over a two-week period. It turned out that 17 of the journals accepted his work for publication as long as he paid them a processing fee of $500 (Segran, 2015). Therefore, young researchers should understand that successfully publishing research is about both publication quantity and publication quality – a goal that cannot be realized through publishing in predatory journals.

Publishing research typically requires properly labeling constructs in writing. For example, using the word mechanism is highly tempting because identifying mechanisms represents a scientific ideal, and papers examining mechanisms empirically and/or theoretically are more likely to be favored (Hommel, 2020). Unlike a pseudo-mechanism, "a mechanism is a structure performing a function in virtue of its component parts, component operations, and their organization. The orchestrated functioning of the mechanism is responsible for one or more phenomena" (Bechtel & Abrahamsen, 2005, p. 423). As such, young scholars should avoid writing about a pseudo-mechanism. For instance, the ability to empathize has been linked to activity in the right temporal–parietal junction, leading to the circular definition of that region as a cortical location related to empathy (Saxe et al., 2004). Researching a true mechanism entails more than just using the term "mechanism" and requires a focus on processes – often over a series of studies.

Relative to writing for the wrong outlet and using the wrong label, the more common problem in writing is a lack of writing. On one end, highly productive writers probably dedicate 30–40 hours a week to writing over the course of many years. In this case, "writing" carries a broad meaning, including activities such as reading the literature one cites and running data analyses, and it is a significant commitment in one's schedule. On the other end, struggling writers go weeks without writing, developing increasingly negative feelings and establishing counterproductive patterns of procrastination and guilt. Addressing these issues early on is necessary for one to complete a manuscript and have a career that depends on effective writing habits. A good resource for understanding and overcoming procrastination is *The Now Habit* (Fiore, 2007) and *How to Change* (Milkman, 2021). See also Chapter 8 in this volume.

Closing Remarks

In previous sections, we discussed important factors that contribute to research productivity. We also provided caveats by listing pitfalls that can jeopardize a scientific career and diminish the reputation of science. All these discussions are based on certain assumptions, such as that the researcher is writing a scientific article rather than a book or a book chapter. We would like to conclude by sharing some general thoughts regarding publishing research.

First, although the rule in publishing is to say something new (Klingner et al., 2005), such a rule is becoming an exception in a publish-or-perish culture. Specifically, the publish-or-perish culture in academia has been described as causing faculty members to decrease publication quality, deemphasize teaching, and experience high levels of stress (Miller et al., 2011). In addition, the culture has caused an excess of publications, many of which are not read or cited (Cano, 2021). The concern over these findings has led to proposals to allow researchers to publish only one paper a year (Nelson et al., 2012). Although we share similar concerns over the publish-or-perish reality, improving academic publishing also demands policy changes from departments, universities, institutions, and grant agencies. Specifically, a balance must be explored between research support and research incentives (Franzoni et al., 2011). Whereas too little support and too strong incentives may reinforce the publish-or-perish culture, the opposite situation may result in a waste of resources. Therefore, stakeholders should consider clarifying research quality expectations and also adjusting research incentive programs (Fanelli et al., 2015).

Second, there is a pervasive gender gap in academic publications. Relative to male researchers, female scholars are less likely to receive tenure and full professorship (Fox, 2005). The gap may result from several factors, including sex segregation in academics, gender differences in productivity, and gender inequality at the time of promotion decision. Evidence suggests that none of the factors can be ruled out in explaining the gender gap (Weisshaar, 2017). In addition, the gender gap in academia is also evident through the citation norms (Wang et al., 2020). A recent *Nature Neuroscience* study investigated 31,418 articles (with 303,886 citations in total) published in the top five neuroscience journals between 2009 and 2018. Their analysis revealed that the publications with females listed as first and last authors were cited 13.9% less, a result based on generalized regression models that controlled for (a) the proportion of different authorship patterns based on the gender of first and last author, (b) year of publication, (c) journal, (d) number of authors, (e) empirical or review research, and (f) seniority of the first and last author (Dworkin et al., 2020).

To call attention to this gender gap in citations, the *Journal of Cognitive Neuroscience* has started to include a “diversity in citation practices” section in all articles since 2021 and explicitly urges authors to consider this factor in their citations. To check the gender balance in article references, reviewers and authors can use web-based tools, such as the Gender Citation Balance Index (GCBI) and Gender Balance Assessment Tool (GBAT). We, thus, encourage junior researchers to remain vigilant with respect to gender inequalities and, by extension, inequalities that disadvantage other minority groups in the process of writing and publishing research (e.g., ethnic minorities and individuals with disabilities).

Third, as one’s scientific career rises with the number of publications, invitations to serve as a reviewer or editor for journals will arrive. Serving in such roles for scientific journals will help a scholar to gain different perspectives of the publication process, and these insights may contribute to publishing one’s own research. In addition, the current trend of science is to move towards an open-science model by making research data, materials, and even peer reviews publicly available (Polka et al., 2018; Wicherts & Bakker, 2012). With such openness, one can learn a lot from reviewing the details of research conducted by others and, if serving as a journal reviewer, should aim to present thorough critiques in a constructive manner (see Chapter 33 in this volume). Finally, it is a good habit to regularly search and read editorial comments and announcements from journals of interest, especially editorials from incoming editors. From such reading, one can usually obtain useful information regarding preferences on topics and methodological approaches (Fowler, 1993). It is also helpful to know the composition of editorial board members of journals in which one aspires to publish. These members can be potential editors and reviewers of one’s submissions to the journal and it never hurts to understand their research perspectives.

Prior to concluding the chapter, we would like to share further resources related to successfully publishing research. For generating research questions, we recommend hearing the advice from McGuire (1997) and Davis (1971), as their work has been listed as “must-read” in some social science laboratories (see also Chapter 3 in this volume). For a more complete list of research reporting guidelines, interested readers can explore this online resource (www.equator-network.org). To improve scientific writing in English, one may find valuable guidance from some helpful books (Gastel, 2016; Greene, 2013) and online resources from universities, such as well-designed courses at the University of Chicago (McEnernehy, 2013) and Harvard University (Carson et al., 2012). For conducting qualitative research, a good resource is the *Handbook of Qualitative Research* (Denzin & Lincoln, 2011; see Chapter 20 in this volume).

Finally, we encourage junior scholars to learn about the history of science or at least the history of their own discipline. One can learn many lessons from reading about the successes of renowned scientists (Weiner, 2003). Another benefit of learning science history is that one can better appreciate the value of one’s own work as part of the history of one's field. Such satisfaction can be critical because, as a scholar, one's friends and relatives may not fully understand one's work, and one is unlikely to get rich as a scientist (Weinberg, 2003). One final benefit from learning about history is that one can gain knowledge about the development of scientific organizations and journals, information that can guide decisions in the publication process (VandenBos, 2017).

References

Aczel, B., Szaszi, B., Sarafoglou, A., Kekecs, et al. (2020). A consensus-based transparency checklist. *Nature Human Behaviour*, *4*(1), 4–6.

American Psychological Association (2017). Ethical principles of psychologists and code of conduct. Available at: www.apa.org/ethics/code/.

Awrey, J., Inaba, K., Barmparas, G., et al. (2011). Reference accuracy in the general surgery literature. *World Journal of Surgery*, *35*(3), 475–479.

Bechtel, W. & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Science Part C*, *36*, 421–441. https://doi.org/10.1016/j.shpsc.2005.03.010

Bem, D. J. (1995). Writing a review article for *Psychological Bulletin*. *Psychological Bulletin*, *118*(2), 172–177. https://doi.org/10.1037/0033-2909.118.2.172

Bem, D. J. (1987). Writing the empirical journal article. In M. P. Zanna & J. M. Darley (eds.), *The Compleat Academic* (pp. 171–201). Lawrence Erlbaum Associates.

Boice, R. & Jones, F. (1984). Why academicians don't write. *Journal of Higher Education*, *55* (5), 567–582.

Bornmann, L. & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the American Society for Information Science and Technology*, *66*(11), 2215–2222. https://doi.org/10.1002/asi

Bossuyt, P., Reitsma, J., & Bruns, D. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. *Annals of Internal Medicine*, *138*, 40–44.

Cano, A. F. (2021). Letter to the Editor : publish, publish . . . cursed ! *Scientometrics*, 126(4), 3673–3682.

Carlson, R. V., Boyd, K. M., & Webb, D. J. (2004). The revision of the Declaration of Helsinki: Past, present and future. *British Journal of Clinical Pharmacology*, *57*(6), 695–713.

Carson, S., Fama, J., Clancy, K., Ebert, J., & Tierney, A. (2012). *Writing for Psychology: A Guide for Psychology Concentrators*. Harvard University.

Chamberlin, T. C. (1897). Studies for students: The method of multiple working hypotheses. *Journal of Geology*, *5*(8), 837–848. https://doi.org/10.3109/13590849208997976

Davis, M. S. (1971). That's interesting! Towards a phenomenology of sociology and a sociology of phenomenology. Philosophy of the social sciences. *Philosophy of the Social Sciences*, *1*(2), 309–344.

Denzin, N. & Lincoln, Y. (2011). *Handbook of Qualitative Research*. SAGE Publications.

Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: A neurophysiological study. *Experimental Brain Research*, *91*(1), 176–180.

Dwan, K., Altman, D. G., Arnaiz, J. A., et al. (2008). Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS ONE*, *3*(8), e3081.

Dworkin, J. D., Linn, K. A., Teich, E. G., et al. (2020). The extent and drivers of gender imbalance in neuroscience reference lists. *Nature Neuroscience*, *23*(8), 918–926.

Etter, J. & Stapleton, J. (2009). Citations to trials of nicotine replacement therapy were biased toward positive results and high-impact-factor journals. *Journal of Clinical Epidemiology*, *62*, 831–837.

Fanelli, D., Costas, R., & Larivière, V. (2015). Misconduct policies, academic culture and career stage, not gender or pressures to publish, affect scientific integrity. *PLoS ONE*, *10*(6), 1–18. https://doi.org/10.1371/journal.pone.0127556

Fiore, N. A. (2007). *The Now Habit: A Strategic Program for Overcoming Procrastination and Enjoying Guilt-Free Play*. Penguin.

Fisher, B. S., Cobane, C. T., Vander Ven, T. M., & Cullen, F. T. (1998). How many authors does it take to publish an article? Trends and patterns in political science. *PS – Political Science and Politics*, *31*(4), 847–856. https://doi.org/10.1017/S1049096500053452

Fiske, D. W. & Fogg, L. (1992). But the reviewers are making different criticisms of my paper! Diversity and uniqueness in reviewer comments. In A. E. Kazdin (ed.), *Methodological Issues & Strategies in Clinical Research* (pp. 723–738). American Psychological Association.

Fowler, R. D. (1993). Statement of editorial policy: 1993. *American Psychologist*, *48*, 5–7.

Fox, M. F. (2005). Gender, family characteristics, and publication productivity among scientists. *Social Studies of Science*, *35*(1), 131–150.

Franzoni, C., Scellato, G., & Stephan, P. (2011). Changing incentives to publish. *Science*, *333* (6043), 702–703. https://doi.org/10.1126/science.1197286

Gastel, B. (2016). *How to Write and Publish a Scientific Paper*. Greenwood.

Grant, L. & Ward, K. B. (1991). Gender and publishing in sociology. *Gender & Society*, *5*(2), 207–223.

Greene, A. E. (2013). *Writing Science in Plain English*. University of Chicago Press.

Haig, B. D. (2005). An abductive theory of scientific method. *Psychological Methods*, *10*(4), 371–388. https://doi.org/10.1037/1082-989X.10.4.371

Hollenbeck, J. R. & Wright, P. M. (2017). HARKing, SHARKing, and THARKing: Making the case for post hoc analysis of scientific data. *Journal of Management*, *43*(1), 5–18. https://doi.org/10.1177/0149206316679487

Hommel, B. (2020). Pseudo-mechanistic explanations in psychology and cognitive neuroscience. *Topics in Cognitive Science*, *12*(4), 1294–1305. https://doi.org/10.1111/tops.12448

Hudson, J. (1996). Trends in multi-authored papers in economics. *Journal of Economic Perspectives*, *10*(3), 153–158.

Ilie, A. & Ispas, D. (2007). Excellence through diversity: Interview with a prolific researcher. *Europe's Journal of Psychology*, *3*(3).

International Committee of Medical Journal Editors (2018). Defining the role of authors and contributors. Available at: www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html.

Ioannidis, J. P. A. & Klavans, R. (2018). The scientists who publish a paper every five days. *Nature*, *561*, 167–169.

Johnson, R., Watkinson, A., & Mabe, M. (2018). The STM Report: An overview of scientific and scholarly publishing. *International Association of Scientific, Technical and Medical*

Publishers, *5*, 212. Available at: www.stm-assoc.org/2018_10_04_STM_Report_2018.pdf.

Jung, H. & Albarracin, D. (2021). Concerns for others increases the likelihood of vaccination against influenza and COVID-19 more in sparsely rather than densely populated areas. *Proceedings of the National Academy of Sciences*, *118*(1), e2007538118.

Kazak, A. E. (2018). Editorial: Journal article reporting standards. *American Psychologist*, *73* (1), 1–2. https://doi.org/http://dx.doi.org/10.1037/amp0000263

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Klesky, K. (2015). *The Professor Is In: The Essential Guide to Turning Your Ph.D. into a Job*. Crown Publishing Group.

Klingner, J. K., Scanlon, D., & Pressley, M. (2005). How to publish in scholarly journals. *Educational Researcher*, *34*(8), 14–20. https://doi.org/10.3102/0013189X034008014

Lapidow, A. & Scudder, P. (2019). Shared first authorship. *Journal of the Medical Library Association*, *107*(4), 618–620.

Lee, A. & Boud, D. (2003). Writing groups, change and academic identity: Research development as local practice. *Studies in Higher Education*, *28*(2), 187–200.

Leimu, R. & Koricheva, J. (2005). What determines the citation frequency of ecological papers? *Trends in Ecology & Evolution*, *20*(1), 28–32.

Li, Y. (2019). Mentoring junior scientists for research publication. In *Novice Writers and Scholarly Publication* (pp. 233–250). Palgrave Macmillan.

Linton, J. D., Tierney, R., & Walsh, S. T. (2011). Publish or perish: How are research and reputation related? *Serials Review*, *37*(4), 244–257. https://doi.org/10.1016/j.serrev.2011.09.001

Liu, S. & Tenenbaum, G. (2018). Research methods in sport and exercise psychology. In *Oxford Research Encyclopedia of Psychology*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190236557.013.224

Luck, S. J. (2014). *An Introduction to the Event-Related Potential Technique*, 2nd ed. MIT Press. https://doi.org/10.1118/1.4736938

Masic, I. (2011). How to search, write, prepare and publish the scientific papers in the biomedical journals. *Acta Informatica Medica*, *19*(2), 68–79.

McDougall, W. (2015). *An Introduction to Social Psychology*. Psychology Press.

McEnernehy, L. (2013). The problem of the problem. In *The University of Chicago Writing Program* (pp. 1–30). University of Chicago Press.

McEnernehy, L. (2021). Leadership lab: The craft of writing effectively. Available at: www.youtube.com/watch?v=vtIzMaLkCaM&t=3085s.

McGrail, M. R., Rickard, C. M., & Jones, R. (2006). Publish or perish: A systematic review of interventions to increase academic publication rates. *Higher Education Research and Development*, *25*(1), 19–35. https://doi.org/10.1080/07294360500453053

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

Milkman, K. (2021). *How to Change: The Science of Getting from Where You Are to Where You Want to Be*. Penguin.

Miller, A. N., Taylor, S. G., & Bedeian, A. G. (2011). Publish or perish: Academic life as management faculty live it. *Career Development International*, *16*(5), 422–445. https://doi.org/10.1108/13620431111167751

Moher, D., Schulz, K. F., & Altman, D. G. (2001). The CONSORT statement: Revised recommendations for improving the quality of report of parallel-group randomised trials. *JAMA*, *285*, 1987–1991. https://doi.org/10.1016/j.ijsu.2010.09.006

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Grp, P. (2010). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *International Journal of Surgery*, *8*(5), 336–341. https://doi.org/10.1371/journal.pmed.1000097

Mudrak, B. (2021). What is a megajournal? Available at: www.aje.com/arc/what-is-a-megajournal.

Munafò, M. R., Nosek, B. A., Bishop, D. V., et al. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 1–9.

Murtaugh, P. A. (2002). Journal quality, effect size, and publication bias in meta-analysis. *Ecology*, *83*(4), 1162–1166.

Neill, U. S. (2008). Publish or perish, but at what cost? *Journal of Clinical Investigation*, *118* (7), 2368.

Nelson, L. D., Simmons, J. P., & Simonsohn, U. (2012). Let's publish fewer papers. *Psychological Inquiry*, *23*(3), 291–293. https://doi.org/10.1080/1047840X.2012.705245

Payne, A. A. & Siow, A. (2003). Does federal research funding increase university research output? *Journal of Economic Analysis & Policy*, *3*(1), 1–20.

Phaf, R. H. (2020). Publish less, read more. *Theory and Psychology*, *30*(2), 263–285. https://doi.org/10.1177/0959354319898250

Platt, J. R. (1964). Strong inference. *Science*, *146*(3642), 347–353.

Polka, J. K., Kiley, R., Konforti, B., Stern, B., & Vale, R. D. (2018). Publish peer reviews. *Nature*, *560*(7720), 545–547. https://doi.org/10.1038/d41586-018-06032-w

Qiu, J. (2010). Publish or perish in China. *Nature*, *463*(7278), 142. https://doi.org/10.1038/463142a

Rawat, S. & Meena, S. (2014). Publish or perish: Where are we heading? *Journal of Research in Medical Sciences*, *19*(2), 87–89.

Reich, E. S. (2009). The rise and fall of a physics fraudster. *Physics World*, *22*(5), 24.

Riviera, E. (2013). Scientific communities as autopoietic systems: The reproductive function of citations. *Journal of the American Society for Information Science and Technology*, *64*(7), 1442–1453. https://doi.org/10.1002/asi

Rozin, P. (2009). What kind of empirical research should we publish, fund, and reward?: A different perspective. *Perspectives on Psychological Science*, *4*(4), 435–439. https://doi.org/10.1111/j.1745-6924.2009.01151.x

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701.

Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: Linking developmental psychology and functional neuroimaging. *Annual Review of Psychology*, *55*, 87–124.

Schweitzer, J. C. (1988). Research article productivity by mass communication scholars. *Journalism Quarterly*, *65*(2), 479–484.

Segran, E. (2015). Why a fake article titled "Cuckoo for Cocoa Puffs?" was accepted by 17 medical journals. Available at: www.fastcompany.com/3041493/why-a-fake-article-cuckoo-for-cocoa-puffs-was-accepted-by-17-medical-journals.

Sharman, A. (2015). Where to publish. *Annals of the Royal College of Surgeons of England*, *97*(5), 329–332. https://doi.org/10.1308/rcsann.2015.0003

Shockley, W. (1957). On the statistics of individual variations of productivity in research laboratories. *Proceedings of the IRE*, *45*(3), 279–290. https://doi.org/10.1109/JRPROC.1957.278364

Smith, H. M. & Banks, P. B. (2017). How dangerous conservation ideas can develop through citation errors. *Australian Zoologist*, *38*(3), 408–413.

Sternberg, R. J. (2000). Writing for your referees. In R. J. Sternberg (ed.), *Guide to Publishing in Psychology Journals* (pp. 161–168). Cambridge University Press.

Strunk, W. (2007). *The Elements of Style*. Penguin.

Tramer, M. R. (2013). The Fujii story: A chronicle of naive disbelief. *European Journal of Anaesthesiology*, *30*(5), 195–198.

van Dalen, H. P. & Henkens, K. (2012). Intended and unintended consequences of a publish-or-perish culture: A worldwide survey. *Journal of the American Society for Information Science and Technology*, *67*(3), 1282–1293. https://doi.org/10.1002/asi

VandenBos, G. R. (2017). From print to digital (1985–2015): APA's evolving role in psychological publishing. *American Psychologist*, *72*(8), 837–847. https://doi.org/10.1037/amp0000229

Vetter, P., Vu, D. L., L'Huillier, A. G., et al. (2020). Clinical features of COVID-19. *BMJ*, *369*, m1470.

von Elm, E., Altman, D., Egger, M., et al. (2007). The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. *Lancet*, *370*, 1453–1457.

Wang, X., Dworkin, J., Zhou, D., et al. (2020). Gendered citation practices in the field of communication. *PsyArXiv*. https://doi.org/10.31234/osf.io/ywrcq

Warren, M. G. (2000). Reading reviews, suffering rejection. In R. J. Sternberg (ed.), *Guide to Publishing in Psychology Journals* (pp. 169–186). Cambridge University Press.

Weinberg, S. (2003). Four golden lessons. *Nature*, *426*(6965), 389. https://doi.org/10.1038/426389a

Weiner, I. B. (2003). *Handbook of Psychology, History of Psychology*. John Wiley & Sons.

Weisshaar, K. (2017). Publish and perish? An assessment of gender gaps in promotion to tenure in academia. *Social Forces*, *96*(2), 529–560. https://doi.org/10.1093/sf/sox052

Wicherts, J. M. & Bakker, M. (2012). Publish (your data) or (let the data) perish! Why not publish your data too? *Intelligence*, *40*(2), 73–76. https://doi.org/10.1016/j.intell.2012.01.004

World Medical Association (2008). Declaration of Helsinki. Available at: www.wma.net/what-we-do/medical-ethics/declaration-of-helsinki/doh-oct2008.

31 Presenting Your Research

Kelly Cuccolo

Abstract

Research presentations offer personal, interpersonal, and professional benefits to students and more senior researchers. Through presentations, students gain important skills (e.g., analytic thinking), are able to meet potential mentors and/or employers, and develop their identities as scholars in a given field. Senior researchers may see increases in motivation, productivity, and collaborative opportunities. Various avenues for presenting one's work include institutional based, regional, national, and international conferences. Readers are encouraged to reflect on logistics and personal and professional goals when deciding on which conference is right for them. Descriptions of poster presentations, oral presentations, and job talks are provided. Subsequently, this chapter offers practical guidance on "best practices" for presenting one's research in each respective modality. Readers are encouraged to reflect on the composition of their audience, the goals of their presentation, and the visual organization of material to craft the most effective presentation possible.

Keywords: Research Presentation, Professional Development, Conference Presentation, Poster Presentations, Job Talk

Introduction

There is a breadth of literature detailing the positive outcomes of students' engagement in research experiences (Helm & Bailey, 2013; Hunter et al., 2007; Seymour et al., 2004). The culmination of students' engagement in research is often the presentation of their research projects that have been facilitated by faculty mentors. The presentation of research boasts benefits for students, including increased communication skills and confidence in public speaking (Kneale et al., 2016). For senior researchers, presenting one's research may be viewed as a professional responsibility to the field. Regardless of where you are in your career, presenting your research affords benefits to you and science as a whole. It allows for knowledge to be shared with the scientific community, authors to obtain feedback on novel and developing work, and collaborative relationships to form (Mata et al., 2010; Smith & Rankin, 2002). Researchers may choose to present their research in a variety of formats, including poster presentations, oral presentations, and symposia. Further, the academic job application process frequently includes "job talks," sometimes referred to as "research talks," where researchers are asked to share their research orally. In this chapter, I will generally discuss why you should present your research, avenues and formats for doing so, and then outline best practices for presenting.

Benefits of Presenting

Benefits for Student Presenters

Students may be preparing to present their research for a variety of reasons and with different goals. For example, some coursework may require students to present research to peers, professors, and/or the academic community (Helm & Bailey, 2013). Research requirements within coursework may call for students to communicate methods, results, and implications of notable research in the field. In other cases, the student may be presenting the culmination of work conducted under the supervision of a faculty member. The presentation may occur locally (e.g., a symposium hosted at the student's home institution), regionally (e.g., a regional meeting for the relevant field), nationally (e.g., a national meeting encompassing all regions in a country), or internationally (e.g., an international meeting drawing scholars from multiple countries).

Similarly, advanced degree students may be presenting their research as part of their journey to becoming a professional in their field and may have received varying levels of supervision from faculty mentors. Students who present their research report increases in self-efficacy (Carpi et al., 2016; Quan & Elby, 2016), skill development (e.g., analytical thinking; Bauer & Bennett, 2003; Ishiyama, 2002), an interest in and pursuit of advanced degrees (Carpi et al., 2016; Quan & Elby, 2016), and a sense of professional identity (Carpi et al., 2016). Notably, students who participate in research also show greater gains in critical thinking, information literacy, and writing compared to students who complete other types of high-impact practices (e.g., internships; Gunnels, 2019). The process of conducting and presenting one's own research also assists the student in "becoming" a scholar (Seymour et al., 2004). As such, presenting their work offers personal (e.g., confidence), interpersonal (e.g., networking), and professional (e.g., field specific knowledge) benefits to students (Helm & Bailey, 2013; Lien et al., 2019).

Benefits for Faculty Members

Faculty members can also derive personal, interpersonal, and professional benefits from presenting their own research, as well as from presenting alongside students. For example, they may become research mentors for students and/or junior faculty for logistical or personal reasons, but, regardless, faculty members who mentor student research report interpersonal and personal benefits from doing such. Specifically, they indicate that having students present research results in a feeling of personal accomplishment (Potter et al., 2009), increased motivation to do research, learning from their students, and personal fulfillment (Morrison-Beedy et al., 2001).

Working collaboratively alongside students and colleagues also offers professional benefits as it allows for the delegation of tasks, increases visibility, drives productivity, and facilitates feedback about the project (Morrison-Beedy et al., 2001). This increase in productivity and collaboration can assist faculty in meeting promotion and tenure requirements (Tien, 2007). Although institutions may vary in their promotion and tenure requirements, a strong publication record of peer-reviewed articles is generally important (Schimanski & Alperin, 2018). For those outside of academia, increases in productivity and collaboration can result in greater efficiency, job security and promotion, and organizational profit (Fulford & Standing, 2014).

Citations are one way in which publications may be further evaluated (Alperin et al., 2019). As such, faculty members may be able to boost their impact through presenting their work at conferences (de Leon & McQuillin, 2020). For example, de Leon & McQuillin (2020) note research that was not presented due to a conference cancellation was significantly less likely to be cited. Further, the positive impact of research visibility differs by career stage. For early career researchers, presentations allow for "maturation," incorporation of feedback, and advancement of future work; for later-stage and more prominent researchers, conference presentations ensure research is "advertised." The diversity of collaborations fostered by conference attendance can often result in work that is novel and of high quality – positively impacting one's career (Catalini et al., 2020).

Overall, presenting ones' research allows the researcher to obtain feedback on their work and become more integrated into the scientific community; this fosters the development of interpersonal, professional, and academic skills (Kneale et al., 2016). Additionally, in the case of faculty members presenting alongside students, students receive benefits such as personal and professional development while faculty mentors may find the experience personally fulfilling.

Avenues for Presenting Your Research

In terms of avenues for presenting ones' work, options range from university research symposia and showcases to international conferences hosted by prominent associations in one's field. Broadly speaking, conferences, regardless of type, afford participants opportunities to network and interact with colleagues from around the world, learn about emerging research and relevant topics, and develop new skills. Goals and logistical constraints can help researchers determine the most appropriate avenue for their presentations.

Institution-Based Research Conferences

Some institutions hold institutional (or departmental) sponsored research conferences that emphasize student participation. For example, the University of Kentucky provides students across disciplines an opportunity to come together and share their scholarly projects; alternatively, some institutions hold scholarship events focused around specific disciplines such as the University of Florida Genetics Institute Graduate Program Showcase.

Institutional research conferences expose researchers to a diversity of topics, methods, and perspectives that may not have been covered in coursework or previous research experiences (Carsrud et al., 1984); they foster connection with other students and faculty members who have similar interests (Caprio & Hackey, 2014), promote professional socialization (Caprio & Hackey, 2014), and provide employer/graduate-school skill development (e.g., communication; Carsrud et al., 1984). Through close and collaborative interactions with students, faculty mentors can delegate tasks and complete projects with increased efficiency (Lei & Chuang, 2009; Morrison-Beedy et al., 2001), gain motivation, grow both professionally and intellectually, derive personal satisfaction and feelings of accomplishment (Buddie & Collins, 2011), and ultimately feel they have an improved standing within the university (Potter et al., 2009). Finally, these conferences provide faculty and students with an opportunity to publish their work (e.g., Digital Commons) making it accessible to a wide audience (Caprio & Hackey, 2014).

Regional Conferences

Regional conferences, hosted in a defined geographical region, tend to attract researchers from one specific regional area and often focus on a particular topic (Davis & Smith, 1992). Regional conferences, usually smaller in size than national or international conferences, provide attendees with the opportunity to network and meet others with similar interests and goals. The forums and workshops offered by these conferences also promote professional development by engaging participants in meaningful discussions about various topics in the field. Finally, the social elements of regional conferences (e.g., dinners) allow participants to continue to establish professional connections and build friendships while gaining insight into the culture of the host location.

The smaller size, welcoming environment, and general orientation towards mentorship may be ideal for students because such environments provide students with feelings of competency and professionalism, while also providing them with ample feedback on their work (see Gumbhir, 2014). Indeed, students who presented research at a regional conference reported increases in self-efficacy and motivation (Helm & Bailey, 2013). For faculty members as well as junior and senior researchers, actively participating in the regional chapter of a national association can also be beneficial for furthering one's career (Thomas et al., 2013). Those active in their local chapters report this participation as being important for their career development, citing service, continuing education, professional development, and networking as benefits (Thomas et al., 2013).

The Eastern Psychological Association, a subregion of the American Psychological Association, holds an annual regional conference. The conference takes place somewhere along the eastern United States (e.g., Boston, Philadelphia, New York City) and promotes a student-friendly environment. For example, this conference has an undergraduate poster session that affords students the opportunity to present their work. This may include results that are non-significant or a replication/presentation of existing work in the field. The conference is also ripe with professional development and networking opportunities.

National Conferences

National conferences are typically hosted by a national professional organization or society and often encompasses smaller regions. For example, the American Psychological Association (APA; www.apa.org) has seven regions (Eastern, Midwestern, New England, Rocky Mountain, Southeastern, Southwestern, and Western), which each hold their own regional conferences but come together for the APA's annual convention. National conferences expose participants to the diversity of the field and provide a unique opportunity to conceptualize one's research in the body of work happening across regions; this may help guide one's future projects (Mata et al., 2010).

There is a plethora of educational opportunities at national conferences that may serve as opportunities for individuals to interact with recognized experts in the field. These educational opportunities may include sessions that count towards required continuing education credits. Continuing education is essential for competency and skill development (Gillies & Pettengill, 1993), represents an important connection to peers and one's field, and can help maintain and improve job satisfaction (Career Professionals of Canada, 2013). In some cases, continuing education may be a professional requirement. For example, completing a certain number of continuing education hours is requirement of licensed psychologists in many countries (Career Professionals of Canada, 2013).

Students may benefit from the opportunity to connect with potential graduate-school mentors, and potential employers (Mata et al., 2010). This may help students develop their applications for coveted positions, as well as assist in identifying desired career paths. Similarly, national conferences often hold spaces for professionals in the field to explore career and employment options. For example, the National Conference on Education (https://nce.aasa.org) has a "Job Central" as part of their conference programming. National conferences also provide participants with opportunities for cultural enrichment – with ample social events focused on the history and culture of the host location (e.g., tours). As such, national conferences present opportunities for education, professional development, and growth.

International Conferences

International conferences are often highly interdisciplinary (or multidisciplinary), showcasing a variety of methods, theories, conceptual frameworks, and representation from a variety of special-interest groups (Association for Information Science and Technology, n.d.; Berchin et al., 2018). Attendance is often motivated by participants’ desire for self-enhancement (e.g., education), conference activities (e.g., interesting programming), and sightseeing (e.g., travel; Rittichainuwat et al., 2001). The sharing of knowledge, experiences, projects, methodology, and initiatives taking place at international conferences has also been noted to be an important driver for promoting issues related to sustainability (Berchin et al., 2018). Indeed, given the scale and complexity of societal issues, it has been argued international collaboration facilitates a holistic view of global challenges by strengthening global ties, allowing experts to share methods, experiences, and best practices, and encouraging the interaction between academic and key stakeholders (Berchin et al., 2018). Participants also have opportunities to become involved with various committees that have a range of responsibilities (e.g., mentoring, developing programming) – representing valuable leadership and networking experience. Overall, international conferences represent an exciting way for participants to engage in professional development, experience the culture of the host location, and develop international collaborative networks.

Deciding on the Conference That Is Right for You

Deciding where to present your research should be a function of your goals for the presentation and any logistical concerns. While institutional research conferences may be required as part of students’ coursework, they are also ideal for senior undergraduate students. Institutional research conferences allow senior students to present their research who might not have had the opportunity to present elsewhere due to time constraints (e.g., relocating for graduate school). These conferences also afford the benefit of costing less than bigger conferences, as travel is usually minimal, and registration costs are low (and sometimes waived).

These conferences are an important educational tool where students can develop communication skills, expand on and apply learned material, gain exposure to diverse methods, theories, and areas of study, develop more collegial relationships with faculty members, and gain confidence (Caprio & Hackey, 2014). As such, institutional research conferences may be ideal for students who are approaching graduation – the timeline from submitting a proposal to presenting is fairly short, students looking to gain marketable job skills and confidence in public speaking, and those with budgetary constraints. For faculty members, institutional research conferences are an ideal way to see skill development and progress among students they have mentored, can be extremely rewarding, provide opportunities to recruit new students into their research labs (advertising), and to complete projects with increased efficiency – including gaining momentum for publication efforts (Kent et al., 2019). These conferences also can afford presenters publication opportunities (Caprio & Hackey, 2014).

If one’s goals are more aligned with professional development, and multi- or interdisciplinary networking, a national or international conference may be more appropriate than a regional or institutional research conference. Indeed, institutional conferences offer many benefits but are limited in size, and diversity of attendees – participants are often from one institution or institutions within a small defined region (e.g., the Red River Valley conference; www.mnstate.edu/academics/majors/psychology/conference). Thus, participants are somewhat limited in the exposure they gain to emerging trends within their own fields, as well as professional development and continuing education opportunities.

National and international conferences have the draw of being hubs for professional development, networking, and exposing participants to the diversity of work being done in a given area. Indeed, students may be able to connect with potential mentors or labs they wish to join, and that may be helpful for preparing applications for graduate school or post-graduation employment (Mata et al., 2010). Faculty and professionals have a wide breadth of continuing education opportunities to choose from that may help them feel connected to the field; additionally, the diversity of attendees allows individuals to hear new perspectives, gain exposure to new methods and trends in the field, and develop skills, knowledge, and relationships to collaboratively address global problems (Berchin et al., 2018). Notably, national and international conferences often change the host location annually, allowing participants to travel and participate in culturally enriching social events.

Both national and international conferences can be expensive, however, as more extensive travel is often necessary, and registration costs are usually higher than regional conferences. The top three barriers to participating in an international conference are time, money, and travel distance (Rittichainuwat et al., 2001). As such, when deciding to present at a national or international conference one should consider the funds they have available for travel and the flexibility of current schedules.

Types of Research Presentations

Common means for presenting one's research include conference poster presentations, oral presentations, and job talks.

Poster Presentations

Posters are popular visual paper presentations where presenters use a visual medium to present content that would be found in a traditional manuscript or paper (Halligan, 2008; Miller et al., 2007). Presenters are typically assigned a specific time and location to present, with expectations that conference attendees with similar interests will engage the presenter in a discussion about the research (Halligan, 2008). The biggest benefit of poster presentations is the ability to interact with the audience – meeting others who are interested in the topic, engaging in discussion about the research, and receiving feedback "in real time" (Everson, n.d; Ilic & Rowe, 2013; Wipke-Tevis et al., 2002). Furthermore, the brief and informal discussions allow the details of the research project to be broken down and discussed; consequently, posters are often perceived as a less intimidating forum for interactions to occur

(Halligan, 2008). Poster presentations, therefore, may be ideal for the novice presenter and students (Everson, n.d.).

Posters stimulate learning and foster many sought-after skills (e.g., written, verbal, and visual communication; Conyers, 2003). Because constructing a poster requires the presenter both to understand a specific content area and reinterpret and organize the information so it reaches an intended audience, presenters further develop skills related to critical thinking, information retrieval, creativity, analysis, and problem solving (Conyers, 2003; Halligan, 2008). Posters may also serve as useful practice in quickly and clearly explaining the importance and implications of one's project – skills that can be applied to other types of presentations (Miller et al., 2007). Given the interactive nature of posters, they are often great networking opportunities (Ilic & Rowe, 2013), thus benefiting both the presenter and attendee. Posters also allow for the research to be showcased to many conference attendees, increasing the exposure that the project receives and the diversity of feedback the presenter gets (Ilic & Rowe, 2013).

In sum, poster presentations are a popular method of displaying information across various conference settings (Halligan, 2008). Since they encourage dialogue between the presenter and audience, poster presentations are often ideal for those looking to receive feedback on their work and network (Wipke-Tevis et al., 2002). They may also be a good choice for those nervous about giving a talk or those who are newer to conferences (Crooks & Kilpatrick, 1998).

Oral Presentations

An oral presentation ("talk") typically involves the presenter standing in front of an audience, telling them about a particular research topic. Lasting approximately 10–30 minutes, presenters typically use a visual aid (e.g., a slide show) and reserve time at the end of the presentation for questions. Talks offer many benefits for students and faculty members. For instance, talks are associated with full publication of the presented research (Hanchanale et al., 2018). Further, presenting information clearly and answering questions about the material are skills employers find valuable and these are bolstered through talks (Lund, 2013). Learning to give an effective presentation is imperative, as one will often be asked to do such in employment, professional, and academic settings (Adler, 2010; Rowh, 2012). Talks may be ideal for those looking to gain skills and confidence in verbal communication and to feel increasingly connected to their professional field (Lund, 2013).

Job Talks

The colloquium, "job talk," or "research talk" is a critical step in the application process for faculty positions in academia (Sura et al., 2019). With some claiming it is "the most important talk you'll ever give" (Ruthig, 2021, oral communication), talks can range in length (~25–60 minutes). Presenters should, however, aim to finish approximately 10 minutes early to accommodate technical difficulties, audience engagement, and questions (Durvasula & Regan, 2006). One aim of a job talk is to make one's research findings accessible to the audience, having them walk away with a “take-home message,” but job talks must also do more – the candidate must situate themselves as a colleague in the department, demonstrate their vision for their research (i.e., a research agenda), speak to their teaching abilities, and convey enthusiasm (Boysen et al., 2018; Mascarelli, 2014).

Your audience for the job talk will likely either be mostly faculty or an equal mix of both students and faculty, depending on the type of institution you are speaking at (Boysen et al., 2018). You can assume members of the search committee will attend your talk, along with other faculty members in your discipline and perhaps those outside of your discipline. Because everyone is likely to have a different research background, focus, experience, and methodological training, it is important that your talk has enough depth to excite those who do similar work but does not isolate those who do not (Aguilar, 2018). The significance or impact of your research needs to be clear (Mascarelli, 2014). Despite the importance of a job talk, many candidates do not receive formal training in this area (Sura et al., 2019).

Best Practices for Poster Presentations

Posters allow many people to learn about your research and can afford presenters important networking opportunities since posters facilitate informal discussions with audience members (Everson, n.d.). As such, an effective poster will facilitate a discussion about your work by conveying what you did and why it matters to a wide audience (Price, 2011). Your goal is to get your main points across to as many people as possible (Hess et al., 2013). Before starting the process, presenters should reflect on what their take home message is – what do you want your audience to learn (Hess et al., 2013)? You may additionally reflect on what the goal of your presentation is and what aspects of the study you want to convey to the audience (Wipke-Tevis et al., 2002). Importantly, identify a main message and keep the poster focused on that message (Hess et al., 2013). One strategy could be to build the message around an important piece of data (Wipke-Tevis et al., 2002).

After coming up with the main message, the presenter is tasked with figuring out how to convey this message to their audience. A good first step is recognizing who your audience is, asking yourself, “who attends this conference and how can I make my message accessible to them?” There will likely be three types of audience members: people in your area familiar with your specialty, people in your field who have different sub-specialties, and people outside of your field (Woolsey, 1989). To appeal to a diverse audience, your poster should give context (e.g., why is the problem important to address?), minimalize jargon and acronyms, and provide an interpretation of results – how does your work make progress on the problem you have identified (Hess et al., 2013). Because looking at a poster is fairly passive, an active component can help promote reciprocal dialogue and can be effective in transferring knowledge (Ilic & Rowe, 2013). Thus, presenters should offer to quickly summarize their research when someone pauses at the poster (Price, 2011).

Relatedly, presenters may want to have mini printout versions of their poster, with their contact information, to hand out to interested parties (Miller, 2007). This will allow for audience members to review the material more thoroughly later and to contact the presenter with questions or potential collaborations (Miller, 2007). Finally, posters take time and should undergo revisions after feedback is received (Everson, n.d.,).

Elements of the Poster

Presenters should begin by planning their poster around their take home message(s), and the poster should be designed to emphasize that message (Miller et al., 2007). The sections of a poster typically mimic that of a manuscript. These include the title, authors and affiliation, objective/aim, background, methods and data, results, conclusions, and implications (Miller et al., 2007; Wipke-Tevis et al., 2002). Poster presentations are typically designed using PowerPoint or similar software (free options include Google Slides, Canva, and PiktoChart) because such programs allow the presenter to easily align and arrange content, see the contrast between color components, and produce and insert illustrative material (e.g., graphs; Marek, et al., 2002). Additionally, it is easy for presenters to print their poster as slides – allowing them to have printed letter-size versions of their poster readily available for interested audience members (Marek, et al., 2002).

When designing a poster, you want to ensure it holds the audience's attention while communicating the intended message. Knowledge will be transferred more efficiently when the presenter hones in on the take-home message that they are attempting to relay (Shilling & Ballard, 2020). As such, keeping the poster concise and allowing for sufficient white space is important; be mindful that the typical dimensions for a poster are 48″ by 36″ (Everson, n.d.; Wipke-Tevis et al., 2002). To ensure concision and enhance readability, place a heading above each individual section of the poster, use bullet points, and avoid long paragraphs (Wipke-Tevis et al., 2002). In terms of overall design, presenters should aim to keep elements of their poster lined up as if working on a grid, as this organization can be used to guide the reader's attention (Everson, n.d.,). Gundogan et al. (2016) recommend that the poster should start with aims and objectives and flow downwards in column format to methods, data, results, and conclusions/implications.

In the "better poster" design, Mike Morrison emphasizes the necessity of reducing clutter and focusing on the critical points to efficiently promote knowledge transfer (Shilling & Ballard, 2020). Morrison argues that typical posters contain too much text that causes the audience to either keep moving, miss critical information, or spend too much time trying to figure out what the poster is about (Greenfield Boyce, 2019). As such, Morrison proposes a poster format that is clean, with the main finding placed directly in the middle of the poster in a large font and plain language (Greenfield Boyce, 2019). This design also includes a QR code that participants can scan with their phones to be taken to a paper or website that contains details of the research (Greenfield Boyce, 2019). The APA has even created their own template for the "better poster," stating this format affords several benefits including fostering conversation, ease of identifying critical takeaways, visual appeal, and encouraging the presenter to be creative, translate their research findings, and focus on important points (Shilling & Ballard, 2020). Visit https://osf.io/qkf3c/ for an example of this format.

Title, Authors, and Affiliation

The title will often be the first thing the audience notices and, thus, should be easy to observe from a distance (40-point type is recommended; see Figure 31.1); tell the audience what your poster is about (Price, 2011), and state any interesting results (Hess et al., 2013; Miller, 2007). For example, “Intermittent Fasters Exhibit Elevated Eating Disorder Symptomatology Compared to Community Norms” is a more effective title than “The Association Between Intermittent Fasting and Eating Disorder Symptomatology.” Presenters are also advised to stay away from word art (e.g., text with reflection or shadow; Everson, n.d.). Presenters may also wish to use their school’s logo near the title to provide a visual representation of their affiliation. In doing so, presenters are advised to save such images to their desktop and use the “import” or “place tool” to insert the image rather than copying and pasting, as the image could become distorted during printing (Everson, n.d.,).

Objectives/Aims and Background

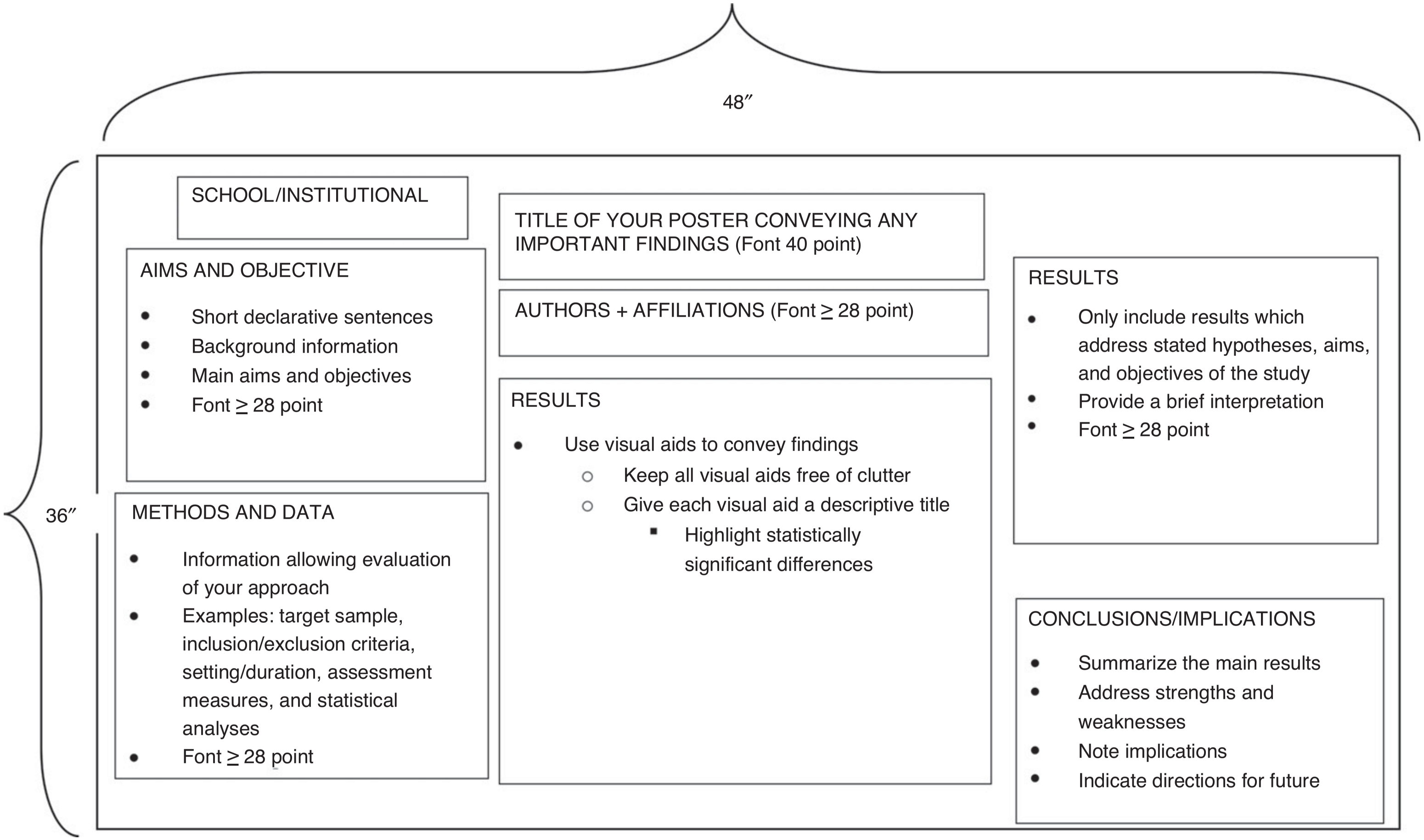
Each section should have a heading, and once you’ve chosen a font for your first section, you should use the same font for subsequent sections. A complimentary font should be used for the body (Everson, n.d.). A good rule of thumb is to use a font size of at least 28 for the body of text. In terms of content in the body of your poster, simple messages tend to be more memorable, so presenters are advised to use short declarative sentences (Price, 2011). This section should contain background information that sets the context, focusing on the main aims and objectives of the research (Gundogan et al., 2016).

Methods and Data

In this section, your goal should be to provide enough information for the audience to evaluate your approach and methodology and situate it within the broader area of study (Miller et al., 2007). This may include the target sample, inclusion/exclusion criteria, setting and duration of the study, assessments and outcome measures, and statistical analyses (Gundogan et al., 2016). Handouts are a useful way to provide interested members of the audience with more detail on your methods and relevant citations (Miller et al., 2007).

Results

Your goal is to present your analyses in a way that addresses the problem or topic your research focused on; instead of making the audience translate your statistical analyses, match your analyses to the questions and concerns of the audience (Miller et al., 2007). Presenters should use visual aids (e.g., graphs), instead of wordy descriptions, to clearly highlight key findings. Let visual aids “speak for themselves” by avoiding chart junk (i.e., non-essential elements; see Marek et al., 2002), giving



SCHOOL/INSTITUTIONAL

TITLE OF YOUR POSTER CONVEYING ANY IMPORTANT FINDINGS (Font 40 point)

AUTHORS + AFFILIATIONS (Font ≥ 28 point)

AIMS AND OBJECTIVE

- Short declarative sentences
- Background information
- Main aims and objectives
- Font ≥ 28 point

METHODS AND DATA

- Information allowing evaluation of your approach
- Examples: target sample, inclusion/exclusion criteria, setting/duration, assessment measures, and statistical analyses
- Font ≥ 28 point

RESULTS

- Use visual aids to convey findings
  - Keep all visual aids free of clutter
  - Give each visual aid a descriptive title
    - Highlight statistically significant differences

RESULTS

- Only include results which address stated hypotheses, aims, and objectives of the study
- Provide a brief interpretation
- Font ≥ 28 point

CONCLUSIONS/IMPLICATIONS

- Summarize the main results
- Address strengths and weaknesses
- Note implications
- Indicate directions for future

Figure 31.1 *Elements of a standard poster.*

each chart an appropriately descriptive title, highlighting statistically significant differences (e.g., with an asterisk or similar symbol), and using the appropriate graph for the data being presented (see Miller, 2007; Wipke-Tevis et al., 2002). Depending on discipline, you may consider incorporating a way to show variability in the sample (e.g., error bars; Wipke-Tevis et al., 2002). The graphs should be large enough for the audience to clearly see (Gundogan et al., 2016). Keep in mind that people who are colorblind will likely have a hard time differentiating red and green (Gundogan et al., 2016)

Of note, the results section should only include results that address stated hypotheses, aims, and objectives of the study. Finally, it is important for presenters to remember that a succinct annotation of the patterns/trends in the chart can go a long way and that findings can be further broken down in the presenter's accompanying project summary (Miller, 2007).

Conclusions and Implications

The conclusions and implications section of the poster should focus on interpreting your findings so that all members of the audience can understand how your work adds to the field while also acknowledging limitations and confounds (Gundogan et al., 2016; Hess et al., 2013). Here, presenters may consider having four sections. Section one summarizes the main results, linking them back to the study's original aims and hypotheses; section two addresses the strengths and weaknesses of the interpretations of the findings; section three notes implications that are tailored to the conference audience (e.g., clinicians, researchers); section four indicates the direction for future research. The length of the conclusions and implications should be roughly equivalent to your other sections (Miller et al., 2007).

Travel and the Day of Presentation

Give yourself ample time to print the poster before transporting it to the location of your presentation. Poster printing may be offered by your institution for free or a fee, or you may need to contact a local retailer (e.g., FedEx). Alternatively, presenters may opt to print their poster "on location," by finding a location close to the conference that provides printing services. Additionally, presenters should prepare single-page poster handouts for interested audience members in accordance with the size of conference attendance (Gundogan et al., 2016). For transport, a mailing tube is the preferred method when using public transportation because you can roll and secure the poster with a rubber band. It is advisable that you take your poster as a carry-on, in case of lost luggage (Utah State University, n.d.).

On the day of presentation, the poster session will typically last about an hour, and one of the presenters should be present at all times to summarize the research and answer questions (Rocky Mountain Psychological Association, n.d.). Presenters may also wish to pack extra pushpins to secure their poster to the poster boards, business cards for networking, and repair supplies in case of a rip (i.e., tape).

Table 31.1 *Tips for an effective poster presentation*

Advised practices	Elements to avoid
Build presentation around a take-home message (Wipke-Tevis et al., 2002)	Isolating audience members with excessive technical details, jargon, and acronyms (Miller, 2007)
Provide context about the purpose, results, and implications of your work (Miller, 2007)	Overwhelming the poster with text or graphics
State your interpretation in the conclusions section (Hess et al., 2013)	Making the font size too small (i.e., < 28 point)
Highlight statistically significant results (Miller, 2007)	Word art, or hard-to-read fonts (Everson, n.d.)
Ensure your title can be read from a distance (40-point type), tells the audience what your poster is about, and states any interesting results (Hess et al., 2013)	Loud colors (e.g., neon green) and/or text that fades into the background (e.g., blue and black)
Give each section an appropriate heading (Everson, n.d.)	Color-blind combinations
Keep fonts consistent and complimentary (Everson, n.d.)	Complicated and "busy" visual aids
Avoid word art (Everson, n.d.)	
Have white space (Everson, n.d.)	
Prepare handouts that can provide the audience with more detailed information (Miller, 2007)	
Short URL to the full paper (Shilling & Ballard, 2020)	

In sum, your poster should be focused on a single message about your research that is accessible to a diverse audience, which can be facilitated by allowing for sufficient white space, thoughtfully constructed visual aids, a succinct interpretation of your findings, and an accompanying oral summary of the poster. See Table 31.1 for a list of effective and ineffective practices. For examples of well-designed posters, see Hess et al., (2013) and Shilling & Ballard (2020).

Best Practices for Oral Presentations

Planning is an essential step, as you want to ensure that your presentation adheres to any specific guidelines set by presentation organizers (Chambers, 2014). Subsequently, you want to consider your audience. Identifying your audience will allow you to be appropriately descriptive and technical. For example, if your audience is familiar with your area of research, you can go into more depth than if your audience contains people from diverse disciplines (Chambers, 2014). As such, ask yourself who the audience is, what background knowledge they have, what they want to know, and what is important to them (Monash University, n.d.). After you have identified your audience and have a sense of the technical requirements of your presentation, you can begin to determine what the goal of your presentation is – what is your take-home message for the audience (Chambers, 2014)? Having a few (~three to five) points you want the audience to walk away with will facilitate concision, and you can consider what information moves you closer to conveying these points to the audience (Adler, 2010; Chambers, 2014).

Slide Organization

It is highly recommended that you plan out the content of your talk before moving into designing the slides. Starting your presentation with an outline is useful because outlines assist in making sure the material flows, helps ensure you are hitting on the point(s) you want your audience to walk away with, and improves learning (Rowh, 2012). One way to organize your slides is to mimic the flow of a manuscript; prime the audience for your talk by introducing the main research question and findings then move into context where you address what has already been done and what your work aims to do. After providing context, you can move into the methods of your research (about one slide). Most of the talk should consist of your results, and you should make sure you summarize the take-home message for the audience. You can end the talk by highlighting the implications and future directions (Adler, 2010).

Slide Design

It is important to design your slides with audience engagement in mind – you want your audience to focus on the content of your talk and avoid having them either stare blankly at your slides or be so focused on reading your slides that they miss the content of your actual speech (Chambers, 2014). Presenters should design each slide using typeset and minimal words (sentence fragments), avoid blocks of text and excessive jargon, and allow sufficient white space (Miller et al., 2007; Rowh, 2012). Miller et al., (2007) recommend no more than six lines of text per slide, keeping text above a font size of 28. Having white space is desirable, as it makes the slides appear clean, and additionally prevents speakers from trying to rush through bulky slides; a good rule of thumb is one minute per slide (Adler, 2010; Chambers, 2014).

Keep in mind that some slides will take more or less time to explain. Slides containing graphs or figures should be thoroughly explained and thus are likely to take more than a minute to discuss (Miller et al., 2007). Given that a typical talk will last between 10 and 30 minutes, presenters should create slides appropriately. Using graphs and figures will help maintain audience engagement (Adler, 2010). The presenter should think about how to accurately represent the data, ensuring that differences between groups, for example, are not misrepresented by restricted or altered axes (Chambers, 2014). Generally, bar graphs (with error bars) may be best suited for comparing groups or variables, line graphs for noting change over time, pie charts for highlighting proportions of a whole, and scatter charts for data that might not follow a trend (Miller et al., 2007).

To this end, graphs and figures should be kept simple and clutter free, and the speaker should always explain what the graph or figure is illustrating in an easy-to-understand manner (Adler, 2010). Relatedly, presenters should be mindful to use a color palette of no more than three colors with a good contrast between the background and the words on the slide. The words should, additionally, be in a font that is big and easy to read; fonts such as Arial, Calibri, and Verdana are easy to read when projected (Chambers, 2014; Rowh, 2012). Presenters should also be cautious about using animations – this may be distracting to the audience (Chambers, 2014). Overall, your slides should complement your talk and "not overpower it" (Kuhn, 2010).

Delivery

Presenters should also *practice* and work to find a presentation style that suits their personality (Adler, 2010). Practicing with diverse types of audience members (friends, mentors, peers, family) will allow you to get a breadth of feedback on your presentation (Adler, 2010) and ensure that your timing is appropriate (Monash University, n. d.). The more you practice, the more comfortable you will be with the content; this will allow you to convey your message to the audience more clearly (Adler, 2010). During your talk, an effective delivery is key, and this includes being aware of your speech and body language. Speak slowly and loudly enough that the audience can follow your key points (Miller et al., 2007). Filler words (e.g., "um") can be distracting, and efforts should be made to discontinue their use (Miller et al., 2007). Make eye contact with your audience often to engage them (Monash University, n.d.).

Finally, come prepared to deal with potential technological glitches and internet connectivity issues. Bring a back-up of your presentation, and do not rely on the internet to access your slides or supplemental materials (American Psychological Association, n.d.). In all, remember that the slides are a tool – they should complement your talk, not steal the show. Given that talks are relatively brief, this type of presentation may be best suited when the researcher has a focused message for the audience; complex work or research with multiple foci may be better suited for a poster presentation (Miller et al., 2007). See Table 31.2 for a breakdown of effective practices.

Best Practices for Job Talks

This section discusses best practices for preparing and delivering a job talk. To deliver a successful job talk, you must put significant effort into planning your presentation. The first step in is to consider your audience (Durvasula & Regan, 2006; Sura et al., 2019); consider the individual members of the audience, the department, and institution you are applying to. In this regard, it may be worthwhile to consider personalizing some of your slides with information specific to the institution at which you are giving the talk (e.g., their mission or

Table 31.2 *Tips for an effective oral presentation*

Advised practices	Elements to avoid
Identify your audience and their background knowledge (Chambers, 2014)	Blocks of text (Rowh, 2012)
Determine what your goal given your audience (Chambers, 2014)	More than six or seven lines of text per slide (Miller, 2007)
Organize your slides to mimic the flow of a manuscript (Adler, 2010)	Excessive jargon (Miller et al., 2007; Rowh, 2012)
Use simple, clean graphs that accurately represent the data (Adler, 2010)	Animations (Chambers, 2014)
Use large fonts that are easy to read (e.g., Calibri; Chambers, 2014)	Talking too fast or quietly (Monash University, n.d.)
Ensure sufficient white space (Rowh, 2012)	Using filler words (Miller et al., 2007)
Use sentence fragments (Miller et al., 2007)	Word art/hard-to-read fonts (Everson, n.d.)
Talk slow and clearly (Miller et al., 2007)	Loud colors (e.g., neon green) and/or text that fades into the background (e.g., blue and black)
Eye contact with the audience (Miller et al., 2007)	Color-blind combinations (e.g., red and green)
Have a backup method of accessing your presentation in case of technological challenges	Complicated and "busy" visual aids

values). This will allow you to situate yourself within their institutional culture and demonstrate that you've considered the "bigger picture." Additionally, consider what background knowledge audience members may have, keeping in mind that members of the search committee, other faculty members in the department, and perhaps those outside of the department will be in attendance. Some useful strategies to assist you include: reading the job ad closely, reading faculty bios and current research, and examining the department and university mission statements (Sura et al., 2019).

Reaching out to the chair of the search committee may also be helpful in tailoring your talk; they will have a better idea of who your audience will be and what the ideal candidate and talk looks like (Sura et al., 2019). Indeed, department chairs often wish more candidates would ask these questions, as it ensures that the candidate is well prepared and that the committee hears a talk that aligns with their expectations (Swalve, 2020, personal communication).

Characteristics of Successful Job Talks

What the department is looking for may vary depending on what type of institution you are speaking at. Committee members from baccalaureate institutions may be focused on evaluating evidence of your teaching ability, and attributes that demonstrate teaching quality are significantly more important among these institutions. In contrast, members from doctoral institutions may be focused on evaluating your research quality, and attributes that demonstrate research quality are significantly more important among these institutions (Boysen et al., 2018).

Relatedly, you want to make sure your talk is accessible to every member of your audience (from students to experts in your field) and highlights your fit within the department and institution. Sura et al., (2019) recommend doing a brief "deep dive" into one technical aspect of your research and then providing a summary of the importance of your findings as they come up – emphasizing their implications for the current research and your own research agenda alongside the reasons why they are exciting. This allows you to impress those in your field while not losing anyone in the audience who is not an expert in your area. You can further accommodate your audience by avoiding excessive jargon and circling back to key themes or messages (Mascarelli, 2014). Indeed, it is critical that you emphasize and remind the audience of the take-home message of your research throughout your talk; having an outline at the beginning of your presentation as well as summary slides built in can help with this (Sura et al., 2019).

Slide Design

Less is more. Avoid complicated slides that contain "add-ons" (e.g., animations), as they can be distracting to the audience, may throw off the flow of your speech, and can make it harder to back-track if an audience member wants you to revisit a specific slide during the questions and answers (Aguilar, 2018). Similarly, text should be minimal because the purpose of the slides is to enhance your presentation. Too much text on the slides can distract the audience and cause them to lose focus on your talk; only put text emphasizing your key points (Sura et al., 2019). Your color scheme should be easy to read and should be accessible to people who are color blind (Gundogan et al., 2016).

Any graphics should be tailored specifically to your talk and not pulled directly from any manuscripts, because you want to provide your audience with clean and less complex visuals that you will have time to explain. If you need to show a complex visual, consider showing it piece by piece so you can explain each aspect as you go. You may also wish to prepare a few slides, placed after the last slide in your talk, that dive into a particular element of your research in more detail, if you anticipate the audience will have questions about it (Sura et al., 2019).

Preparing and Delivering Your Talk

Prior to delivering the talk – practice! The more you practice with colleagues, friends, and family, the more opportunities you have to get a sense of what questions may come up, what works (and what doesn't), check your timing, edit the content and slide design, and get comfortable with the material (Durvasula & Regan, 2006). Being comfortable with your material will go a long way during your actual talk. During the talk, it is advisable to set the stage by introducing yourself and your work to the audience, perhaps focusing on how you became interested in this area, and set some general guidelines (e.g., you will answer questions at the end; Durvasula & Regan, 2006). Everyone has a presentation style that works best for them – do not force something that doesn't work for you (e.g., humor; Durvasula & Regan, 2006).

During the question-and-answer part of your talk, paraphrasing any question you get can be helpful to make sure you fully understand what the audience member is asking, and it will give you more time to think of your response (Durvasula & Regan, 2006). Something along the lines of "What I understand you to be asking is . . .. " or " . . . – is that correct?" (Aguilar, 2018). *If you don't know the answer to a question, be honest,* but *demonstrate your critical thinking skills* by relating it to something you *do* know about or have done (Sura et al., 2019). For example, during one of my job talks, where I presented some of my research on intermittent fasting and eating disorder symptomatology, a student asked how men's motivation for fasting might impact the development of symptomatology. I thanked the student for this thoughtful question, admitting that I did not have a clear answer given the limited research on the etiology of men's eating disorder symptomatology, especially regarding intermittent fasting. However, I linked the question back to the research I was knowledgeable about and told them my hypothesis. The student appreciated my humility and was excited to discuss ways to study this question (and I was later offered this position!). If someone voices a fair critique of your work, then respectfully acknowledge it and use it as an opportunity to discuss how you will address it in future work (Sura et al., 2019). If a question was particularly long or complex, it may also be appropriate to end your response by asking if you answered the question (Aguilar, 2018).

You can deliver an effective job talk by doing work ahead of time to ensure you know who your audience is and what they are looking for. Centering your talk around a take-home message that emphasizes your research findings and their implications can help keep all members of the audience engaged. Make sure you are well practiced so that you feel comfortable with your content during the talk. If you are traveling to give a talk, I echo what I've outlined previously regarding poster presentations – bring everything you need to present as a carry on to circumvent potential issues with lost luggage. See Table 31.3 for a list of effective practices.

Conclusion

There is a plethora of options for presenting one's research, and the manner and context in which a presentation is given should be decided based on a combination of logistic and personal factors. Each type of presentation has its own strengths that the researcher should consider as well. Poster presentations allow the presenter to obtain feedback on their work, in real time, through conversation with audience members. Oral presentations increase one's sense of professionalism, build valued and marketable skills, and are associated with future publication of the presented research. Job talks provide one with the exciting opportunity to share their work with potential future colleagues while also potentially getting a great job.

Table 31.3 *Tips for an effective job talk*

Advised practices	Elements to avoid
Know your audience (Sura et al., 2019).	Animations (Aguilar, 2018)
Consider individual audience members, as well as the department and institution	Word art/hard-to-read fonts (Everson, n.d)
Familiarize yourself with the job advert, search committee, faculty's current research, and department/institutional mission statements (Sura et al., 2019)	Loud colors (e.g., neon green) and/or text that fades into the background (e.g., blue and black) Color-blind combinations (e.g., red and green)
Highlight your fit	Complicated and "busy" visual aids
Have an outline (Sura et al., 2019)	Presentation styles that don't fit your personality (Durvasula & Regan, 2006)
Practice!	
End the talk ~10 minutes early to accommodate technical difficulties and questions	

An effective presentation is one that conveys your message to a diverse audience, and, therefore, spending significant time thinking about the message you want to convey and who your audience is can be instrumental in crafting an effective presentation. You additionally want to ensure that your presentation is visually effective. For posters, this means minimizing text. For talks, this means having slides that complement your speech. Both posters and talks should use clean graphics to showcase results and highlight the key points you want the audience to remember.

References

Adler, A. (2010). Talking the talk: Tips on giving a successful conference presentation. Available at: www.apa.org/science/about/psa/2010/04/presentation.

Aguilar, S. (2018). Tips for a successful job talk. Available at: https://stephenaguilar.com/tips-successful-job-talk/.

Alperin, J. P., Nieves, C. M., Schimanski, L. A., et al. (2019). Meta-research: How significant are the public dimensions of faculty work in review, promotion and tenure documents? *ELife*, *8*, e42254.

American Psychological Association (n.d.) APA 2021 *Presenter Hub*.

American Psychological Association (n.d.). Quite possibly the world's worst PowerPoint presentation ever. Available at: www.apa.org/gradpsych/2012/01/worst-powerpoint-ever.pdf.

Association for Information Science and Technology (n.d.). ASIS&T Annual Meeting. Available at: www.asist.org/meetings-events/am.

Bauer, K. W. & Bennett, J. S. (2003). Alumni perceptions used to assess undergraduate. *Journal of Higher Education*, 74(2), 210–230.

Berchin, I. I., Sima, M., de Lima, M. A., et al. (2018). The importance of international conferences on sustainable development as higher education institutions' strategies to promote sustainability: A case study in Brazil. *Journal of Cleaner Production*, *171*, 756–772. https://doi.org/10.1016/j.jclepro.2017.10.042

Boysen, G. A., Jones, C., Kaltwasser, R., & Thompson, E. (2018). Keys to a successful job talk: Perceptions of psychology faculty. *Teaching of Psychology*, *45*(3), 270–277. https://doi.org/10.1177%2F0098628318779277

Buddie, A. M. & Collins, C. L. (2011). Faculty perceptions of undergraduate research. *PURM: Perspectives on Mentoring Undergraduate Researchers*, *1*(1), 1–21.

Caprio, M. & Hackey, R. (2014). If you built it, they will come: Strategies for developing an undergraduate research conference. *The Journal of Health Administration Education*, *31*(3), 247. Available at: www.proquest.com/scholarly-journals/if-you-built-they-will-come-strategies-developing/docview/1694861709/se-2?accountid=26228.

Carpi, A., Ronan, D. M., Falconer, H. M., & Lents, N. H. (2016). Cultivating minority scientists: Undergraduate research increases self-efficacy and career ambitions for underrepresented students in STEM. *Journal of Research in Science Teaching*, *54* (2), 169–194. https://doi.org/10.1002/tea.21341

Career Professionals of Canada (2013). The benefits of attending professional conferences. Available at: https://careerprocanada.ca/benefits-attending-professional-conferences/.

Carsrud, A. L., Palladino, J. J., Tanke, et al. (1984). Undergraduate psychology research conferences: Goals, policies, and procedures. *Teaching of Psychology*, *11*(3), 141–145.

Catalini, C., Fons-Rosen, C., & Gaulé, P. (2020). How do travel costs shape collaboration? *Management Science*, *66*(8), 3340–3360. https://doi.org/10.1287/mnsc.2019.3381

Chambers, R. (2014). Presenting your research effectively. Available at: www.apa.org/science/about/psa/2014/02/presenting.

Conyers, V. (2003). Posters: An assessment strategy to foster learning in nursing education. Available at: https://doi.org/10.3928/0148-4834-20030101-09.

Crooks, D. & Kilpatrick, M., (1998). In the eye of the beholder: Making the most of poster presentations – Part 2. *Canadian Oncology Nursing Journal* 8 (3), 154–159. https://europepmc.org/article/med/9814152

Davis, S. F. & Smith, R. A. (1992). Regional conferences for teachers and students of psychology. In A. E. Puente, J. R. Matthews, & C. L. Brewer (eds.), *Teaching Psychology in America: A History* (pp. 311–328). American Psychological Association. https://doi.org/10.1037/10120-013

Durvasula, R. S. & Regan, P. C. (2006). Style and substance: Twelve tips for a better job talk. Available at: www.psychologicalscience.org/observer/style-and- substance-twelve-tips-for-a-better-job-talk.

de Leon, F. L. L. & McQuillin, B. (2020). The role of conferences on the pathway to academic impact evidence from a natural experiment. *Journal of Human Resources*, *55*(1), 164–193. http://jhr.uwpress.org/content/55/1/164.short

Everson, K. M. (n.d.). The scientist's guide to poster design. Available at: www.kmeverson.org/academic-poster-design.html.

Fulford, R. & Standing, C. (2014). Construction industry productivity and the potential for collaborative practice. *International Journal of Project Management*, *32*(2), 315–326. https://doi.org/10.1016/j.ijproman.2013.05.007

Gundogan, B., Koshy, K., Kurar, L., & Whitehurst, K. (2016). How to make an academic poster. *Annals of Medicine and Surgery*, *11*, 69–71. https://doi.org/10.1016/j.amsu.2016.09.001

Gillies, D. A. & Pettengill, M. (1993). Retention of continuing education participants. *The Journal of Continuing Education in Nursing*, *24*(1). https://doi.org/10.3928/0022-0124-19930101-06

Greenfield Boyce, N. (2019). To save the science poster, researchers want to kill it and start over. Available at: www.npr.org/sections/health-shots/2019/06/11/729314248/to-save-the-science-poster-researchers-want-to-kill-it-and-start-over.

Gumbhir, V. K. (2014). From students to scholars: Undergraduate research and the importance of regional conferences. *The American Sociologist*, *45*(2), 298–300. https://doi.org/10.1007/s12108-014-9212-2

Gunnels, C. W. (2019). Undergraduate research develops transferable skills more successfully than other high impact practices. Available at: https://stars.library.ucf.edu/research symposium/2019/12thAnnual/3/.

Halligan, P. (2008). Poster presentations: Valuing all forms of evidence. *Nurse Education in Practice*, *8*(1), 41–45. https://doi.org/10.1016/j.nepr.2007.02.005

Hanchanale, S., Kerr, M., Ashwood, P., et al. (2018). Conference presentation in palliative medicine: Predictors of subsequent publication. *BMJ Supportive & Palliative Care*, *8*(1), 73–77. http://dx.doi.org/10.1136/bmjspcare-2017-001425

Helm, H. W. & Bailey, K. G. (2013). Perceived benefits of presenting undergraduate research at a professional conference. *North American Journal of Psychology*, *15*(3). https://digitalcommons.andrews.edu/behavioral-pubs/63/

Hess, G., Tosney, K., & Liegel, L. (2013). Creating effective poster presentations. Available at: https://projects.ncsu.edu/project/posters/.

Hunter, A. B., Laursen, S. L., & Seymour, E. (2007). Becoming a scientist: The role of undergraduate research in students' cognitive, personal, and professional development. *Science Education*, *91*(1), 36–74. https://doi.org/10.1002/sce.20173

Ilic, D. & Rowe, N. (2013). What is the evidence that poster presentations are effective in promoting knowledge transfer? A state of the art review. *Health Information & Libraries Journal*, *30*(1), 4–12. https://doi.org/10.1111/hir.12015

Ishiyama, J. (2002). Does early participation in undergraduate research benefit social science and humanities students? *College Student Journal*, *36*(3), 381–387.

Kent, C., Allen, P. J., Harding, S., & Fielding, J. L. (2019). The Psychology Undergraduate Research Conference: A pathway to publishing?. *Frontiers in Psychology*, *10*, 491. https://doi.org/10.3389/fpsyg.2019.00491

Kneale, P., Edwards-Jones, A., Walkington, H., & Hill, J. (2016). Evaluating undergraduate research conferences as vehicles for novice researcher development. *International Journal for Researcher Development*, *7*(2), 159–177. https://doi.org/10.1108/IJRD-10-2015-0026

Kuhn, J. (2010). 14 Tips for better presentation slides. Available at: www.viget.com/articles/14-tips-for-better-presentation-slides/

Lei, S. A. & Chuang, N. K. (2009). Undergraduate research assistantship: A comparison of benefits and costs from faculty and student perspectives. *Education*, *130*, 232–240.

Lien, A., Fyne, A., DeVito, J., et al. (2019). Promoting undergraduate student engagement in the SCRA Biennial Conference. *Global Journal of Community Psychology Practice*, *10*(2). Available at: www.gjcpp.org/en/article.php?issue=32&article=195.

Lund, N. (2013). Ten years of using presentations at a student conference as a final assessment. *Psychology Learning & Teaching*, *12*(2), 185–188. https://doi.org/10.2304%2Fplat.2013.12.2.185

Mascarelli, A. (2014). Research tools: Jump off the page. *Nature*, *507*(7493), 523–525. https://doi.org/10.1038/nj7493-523a

Marek, P., Christopher, A. N., & Koenig, C. S. (2002). Applying technology to facilitate poster presentations. *Teaching of Psychology*, *29*(1), 70–72. https://doi.org/10.1207%2FS15328023TOP2901_12

Mata, H., Latham, T. P., & Ransome, Y. (2010). Benefits of professional organization membership and participation in national conferences: Considerations for students and new professionals. *Health Promotion Practice*, *11*(4), 450–453. https://doi.org/10.1177%2F1524839910370427

Miller, L., Weaver, A., Johnson, C. (2007). Giving a good scientific presentation. Available at: www.asp.org/education/EffectivePresentations.pdf

Monash University (n.d.). A guide to oral presentations. Available at: –www.monash.edu/rlo/quick-study-guides/a-guide-to-oral-presentations

Morrison-Beedy, D., Aronowitz, T., Dyne, J., & Mkandawire, L. (2001). Mentoring students and junior faculty in faculty research: A win–win scenario. *Journal of Professional Nursing*, *17*(6), 291–296. https://doi.org/10.1053/jpnu.2001.28184

Potter, S. J., Abrams, E., Townson, L., & Williams, J. E. (2009). Mentoring undergraduate researchers: Faculty mentors' perceptions of the challenges and benefits of the research relationship. *Journal of College Teaching & Learning (TLC*), *6*(6). https://doi.org/10.19030/tlc.v6i6.1131

Price, M. (2011). The perfect poster: Experts reveal the art behind displaying your science. Available at: www.apa.org/gradpsych/2011/01/poster.

Quan, G. M. & Elby, A. (2016). Connecting self-efficacy and views about the nature of science in undergraduate research experiences. *Physical Review Physics Education Research*, *12*, 020140 . https://doi.org/10.1103/PhysRevPhysEducRes.12.020140

Rocky Mountain Psychological Association (n.d.). Tips for presenters. Available at: www.rockymountainpsych.com/tips-for-presenters.

Rowh, M. (2012). Power up your PowerPoint: Seven research-backed tips for effective presentations. Available at: www.apa.org/gradpsych/2012/01/presentations

Rittichainuwat, B. N., Beck, J. A., & Lalopa, J. (2001). Understanding motivations, inhibitors, and facilitators of association members in attending international conferences. *Journal of Convention & Exhibition Management*, 3(3), 45–62.

Schimanski, L. A. & Alperin, J. P. (2018). The evaluation of scholarship in academic promotion and tenure processes: Past, present, and future. *F1000Research*, *7*, 1605. https://doi.org/10.12688/f1000research.16493.1

Seymour, E., Hunter, A. B., Laursen, S. L., & DeAntoni, T. (2004). Establishing the benefits of research experiences for undergraduates in the sciences: First findings from a three-year study. *Science Education*, *88*(4), 493–534. https://doi.org/10.1002/sce.10131

Shilling, R. D. and Ballard, D. (2020). Rethinking the science poster. Available at: https://convention.apa.org/blog/rethinking-the-science-poster.

Smith, S. E. & Rankin, C. (2002). *Conferences: Why to Attend and How to Benefit*. *University of Texas at Austin*.

Sura, S. A., Smith, L. L., Ambrose, M. R., et al. (2019). Ten simple rules for giving an effective academic job talk. *PLoS Computational Biology*. *15*(7). https://doi.org/10.1371/journal.pcbi.1007163

Thomas, M., Inniss-Richter, Z., Mata, H., & Cottrell, R. R. (2013). Career development through local chapter involvement: Perspectives from chapter members. *Health*

Promotion Practice, *14*(4), 480–484. https://doi.org/10.1177%2F1524839913479378

Tien, F. F. (2007). Faculty research behaviour and career incentives: The case of Taiwan. *International Journal of Educational Development*, *27*(1), 4-17. https://doi.org/10.1016/j.ijedudev.2006.04.014

Utah State University (n.d.). Poster traveling tips: How far can you go? Available at: https://engineering.usu.edu/students/open-access-computer-labs/poster-traveling-tips.

Wipke-Tevis, D. D., & Williams, D. A. (2002). Preparing and presenting a research poster. *Journal of Vascular Nursing*, *20*(4), 138–142. https://doi.org/10.1067/mvn.2002.130001

Woolsey, J. D. (1989). Combating poster fatigue: how to use visual grammar and analysis to effect better visual communications. *Trends in Neurosciences*, *12*(9), 325–332. https://doi.org/10.1016/0166-2236(89)90039-8

32 Building Fruitful Collaborations

Mary G. Carey and Wendy M. Brunner

Abstract

Collaborating on a scientific endeavor can take extra time, work, and intention to ensure that the collaboration is fruitful. However, it also comes with many benefits, such as the building of professional relationships. There are several best practices that can help increase the likelihood that a collaboration will be successful. These include taking time at the beginning of the collaboration to plan how the team will work together. Teams that are characterized by trust, open communication, and shared goals and expectations, among other qualities, are more likely to be successful. Different forms of interdisciplinary research move researchers from a focus on one's own discipline to increasing integration across other disciplines. Despite the challenges that come with interdisciplinary research, such as navigating differences in discipline-specific practices, such a collaboration can provide the capacity to address scientific problems that are too big for one discipline.

Keywords: Scientific Collaboration, Scientific Productivity, Interdisciplinary Research, Tools of Collaboration, Best Practices for Collaborations, Community-Based Participatory Research

Introduction

Fruitful collaboration is a relationship that produces good and useful results – and is productive (Lindner et al., 2018). The productivity of scientists is generally judged by how much impact they have during a certain time period. Generative products include publications, patents, inventions, presentations, government reports, and product developments. Of note, especially in research-intensive universities, productivity more directly refers to publication generation because most research reports are journal-based publications.

For some time now, research has been moving towards collaboration, across all sciences (National Academy of Sciences et al., 2005; National Research Council, 2014). The investigation and revelation of knowledge is shifting from individual efforts to group work, from single to multiple institutions, and from national to international. Interdisciplinary research is required to address serious global challenges (e.g., infectious disease pandemics) (Moradian et al., 2020). There is a clear expectation, from academic institutions and funding agencies, to collaborate, interface, cooperate, join forces, coproduce, partner, or co-act with one another. Whether in education, nursing, business, psychology, biology, or any of the other dozens of social and behavioral science disciplines, conducting research collaboratively is strongly encouraged both within and across disciplines.

The interdisciplinary process involves collaboration, framing the right question, information searches, and knowing how to utilize that information. Thus, it is particularly well suited to contemporary students. Klein led the early exploration of modern collaborative studies, recognizing both the advantages and barriers (Klein, 1990). Davis wrote an article entitled "Interdisciplinary courses and team teaching: New arrangements for learning," which specifically addressed interdisciplinary, team-teaching best practices (Davis, 1995). In addition, Newell and Klein, Szostak, and Repko are among others who have made considerable contributions to the understanding and application of interdisciplinary research (Newell & Klein, 1996; Repko, 2008; Szostak, 2002). Collaboration and integration, they believe, are the keys to scientific progress. True collaborators are usually academics involved in scholarly activities that surpass the typical disciplinary boundaries; while focusing on research within their respective disciplinary boundaries, they utilize concepts and techniques from other disciplines as well. It is well known that fruitful collaborations improve professional practice and targeted outcomes (e.g., student development) (Bridges et al., 2011; Goldsberry, 2018; Reeves et al., 2017).

What Makes a Collaboration Work?

Successful research collaborations begin with agreement among team members on the research question, objectives and overall approach, core concepts and terminology, roles and responsibilities of the team members, and what the products or outcomes of the research will be. It is important to understand and appreciate the expertise that is needed to address the research question at hand. In addition, teams need to recognize that collaborative work takes more time; it is important to add additional time into research timelines. While much of this chapter will focus on interdisciplinary work, these practices are just as important for research collaborations within disciplines – intradisciplinary research.

Bennett and Gadlin (2012) interviewed National Institutes of Health researchers on teams that did and did not work effectively to identify the characteristics of successful teams. What they found was that the key factor in effective collaborations is trust, which they acknowledge may seem beside the point to researchers focused on their scientific work. Trust underlies the necessary giving up of one's individual control of the research process to one's collaborators. Bennett and Gadlin (2012) recommend that teams actively work to build trust by clearly setting out roles and expectations at the beginning of a collaborative project. "We believe that trust, whether grounded in a strong personal relationship or created and reflected in a written agreement, plays a critical role in the functioning of scientific teams and collaborations" (Bennett and Gadlin, 2012, p. 5)

In addition to trust, they note that successful research teams also have effective leadership, open communication, shared expectations, clear roles and responsibilities, a shared vision, and a process for sharing recognition and credit. Team members take time to learn the terminology of the other researchers. They have self-awareness and awareness of others – understanding themselves in terms of characteristics (e.g., how they communicate and how they deal with conflict) and extending that understanding to others on the team so that all can learn how best to work together. This can be done informally or through formal inventories (e.g., 360-degree assessment, also known as multi-rater feedback; Atkins & Wood, 2002). Effective teams strive to "support the scientific disagreement while containing the personal conflict." They do not avoid conflict – differences are likely to come up, especially in interdisciplinary work – and promote and support respectful dialogue to work through these issues. Because there will be times when teams are not able to resolve conflict through dialogue, Bennett and Gadlin (2012) recommend having a plan with other steps that can be taken, such as bringing in an outside party to mediate the situation.

Finally, Bennett and Gadlin (2012, p.11) observe that members of effective teams enjoy both their scientific work and their work with the team. "We have heard from many researchers that a good collaboration provides many benefits beyond strengthening the actual science. They cite many intangible elements, such as complementarity of work styles and approaches, improved quality of the experimental design or analysis of the results, and strong personal connections to colleagues, which are not merely supportive but also deeply enjoyable and satisfying."

Interdisciplinary Research

Interdisciplinary research is a broad category that includes three distinct subtypes – *multidisciplinary, interdisciplinary, and transdisciplinary* – which each differ according to the problem focus, the nature of the interpersonal relationships, the cognitive processes, and the expected outcomes (including the effects on the disciplines and the degree of integration between disciplines).

Multidisciplinary Research

Multidisciplinary research is considered the most basic subtype of interdisciplinary research. In this approach to team science, members focus on a common problem, with each using and maintaining their disciplinary perspective. For example, a team investigating a new drug may include pharmacists, epidemiologists, biostatisticians, and clinicians, each carrying out an individual role. The problem may originate from a single discipline, but each discipline approaches the problem with a distinct or individual goal (Breckler, 2005; Choi & Pak, 2006). Researchers on a multidisciplinary team work independently on different aspects of the project, either concurrently or sequentially, but the knowledge and methods are pooled to identify effective solutions. Hence, interactions between team members are minimal, and communications from individual investigators are typically with the project leader.

The cognitive process associated with multidisciplinary research is described as "additive" in that distinct perspectives and tools are brought to bear separately. While new perspectives may add depth or breadth to the understanding of the problem, or new methods or approaches to address it, the group members do not engage in the cognitive work to integrate or synthesize the distinct perspectives into a new coherent whole. In the process of multidisciplinary work, members learn about each other; individual knowledge is gained, but existing discipline-specific knowledge is not modified, changed, or challenged. Discipline-specific knowledge or methods may be extended or changed as the result of application to the problem. Purists strive to understand their discipline, its structure, and the quest for continuing to develop knowledge that advances the discipline because it emanates from the discipline's domain and perspective. Thus, discipline-specific knowledge is not changed because, in the strictest sense, cross-disciplinary integration and synthesis does not occur.

The multidisciplinary approach is considered limited because the problem focus is narrow and the outcomes do not include the development of new cross-cutting concepts, models, or approaches (Repko, 2008). For example, imagine having all of the ingredients to bake a cake but not being able mix them together – nothing new is created. However, unlike the more integrated approaches of inter- and transdisciplinary research, multidisciplinary research does provide an opportunity to focus on discipline-specific problems and may serve as a first step in team building and cross-disciplinary idea exchange that lead to fruitful collaborations.

Interdisciplinary Research

Interdisciplinary research is the label used for a specific subtype of team science as well as the broad category label used to refer the total continuum. As a subtype, interdisciplinary research is considered a more robust approach to collaboration and knowledge integration compared to the multidisciplinary approach (Stokols et al., 2008). Problems addressed using the interdisciplinary approach are typically both broad in scope and complexity. Consequently, the knowledge and skills required extends beyond the bounds of a single discipline. With the interdisciplinary approach, team members from different disciplines have shared goals; however, individuals work from their own disciplinary foundation. Importantly, the social and scientific process is interactive. Members work jointly on a project, and collaborative efforts among members are expected and encouraged (e.g., manuscript and grant preparation). The cognitive work of interdisciplinary teams is the exchange and integration of knowledge across disciplinary boundaries, such that an expected outcome of this work is building fruitful collaborations. Members learn about each other's perspective and work to create new knowledge that reflects an integration or synthesis of the parts. For example, Carey and colleagues suggest that a collaborative, interdisciplinary, international scientific team provides better understanding of the prevalence of major psychological distressors among first responders (Carey et al., 2021). The outcomes of this approach are outside the boundaries of any single participating discipline and include new knowledge that is a blending of the

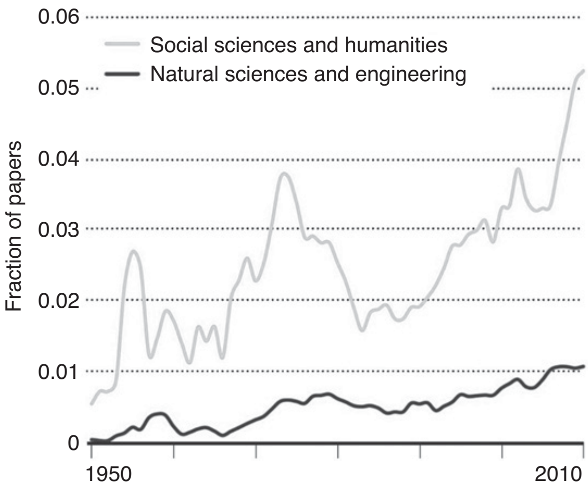


Figure 32.1 *Papers with "interdisciplinary" in title (VanNoorden, 2014).*

discipline-specific parts (Choi & Pak, 2006; Stokols et al., 2008). Thus, team members really value this benefit, and it serves them well throughout their academic career. Figure 32.1 depicts an exponential increase in publications with the word "interdisciplinary" in the title among the social sciences and humanities versus the natural sciences and engineering.

A useful example of an interdisciplinary approach comes from an article about research on electronic cigarettes, which provides evidence to inform decision making related to tobacco product regulation, "Answering questions about electronic cigarettes using a multidisciplinary model." While Breland and colleagues (Breland et al., 2019) characterize their work as multidisciplinary, their descriptions of their research model with team members "extend[ing] their interests beyond the boundaries of their discipline to collaborate effectively with the shared goal of producing the rigorous science needed to inform empirically-based tobacco policy" (Breland et al., 2019, p. 368) indicate a degree of integration more characteristic of interdisciplinary research.

The research team includes members from the fields of psychology, analytical chemistry, aerosol research, biostatistics, engineering, internal medicine, and public health. The team reported that the benefits of this approach included better science and opportunities to build ongoing collaboration. They noted the need to stay focused on original goals, as the interdisciplinary approach had spurred new questions of interest. Challenges of this approach included steep learning curves, as team members learned basic concepts and terminology from the other disciplines. This required careful listening and active engagement on behalf of those new to a particular discipline and clear explanations of basic concepts and terminology by those who were the experts in that area. The authors noted that this intensive learning process can be difficult for those who are experts in their own fields (Breland et al., 2019).

Another example of interdisciplinary approaches to research comes from hazards and disaster research that includes researchers from the fields of engineering, public health, social sciences, natural sciences, risk analysis, and urban planning. In hazards and disaster research, disciplinary boundaries are naturally porous as the disasters themselves reveal the "deep interconnections" between the systems and other factors related to the disaster being studied (Peek & Guikema, 2021).

Transdisciplinary Research

Transdisciplinary research is considered the most complex and challenging subtype of interdisciplinary research. The approach is problem-driven and addresses multiple levels within which the problem is embedded simultaneously. As such, a transdisciplinary team may include scientists whose domains span from the basic cellular science to societal behavioral structural levels. More recent conceptions of transdisciplinary teams include non-scientist stakeholders and may include workers, students, and scientists from relevant disciplines (Choi & Pak, 2006). Several experts note that to conduct transdisciplinary research the team must mature over time, such that mutual trust and respect among members is established.

Choi and Pak (2006) described that members of transdisciplinary teams must be able to both "role release" (e.g., accept that others can do what one as specialist was trained to do) and "role expand" (e.g., engage in work beyond what one was specifically trained to do). The cognitive work includes the disassembling of discipline-specific knowledge and creative reassembly to form a new framework that transcends the discipline-specific models (Choi & Pak, 2006). In this approach, team members are working beyond their disciplines in a new realm that was creatively constructed by the team. Outcomes of this approach include new concepts, theories, approaches, and also new disciplines or fields of study (e.g., psychoneuroimmunology).

An example of the transdisciplinary approach comes from the discovery of "nursing informatics." The Healthcare Information and Management Systems Society defines nursing informatics as a specialty that integrates nursing science, computer science, and information science to manage and communicate data, information, knowledge, and experience in nursing practice (Garcia-Dia, 2021). Thus, nursing informatics was generated as a new field of study with the combination of multiple sciences. A second example comes from the study of genetic and social influences on disease in which researchers from sociology and genetics have integrated theories from their respective disciplines to come up with a new approach to studying disease causality (Pescosolido et al., 2008).

In summary, the three subtypes of interdisciplinary research form a continuum of increasing complexity in goals, interpersonal relationships, cognitive process, outcomes, and integration across disciplines. Figure 32.2 depicts these degrees of integration. However, regardless of approach (i.e., multidisciplinary, interdisciplinary, or transdisciplinary), the fundamental starting point is grounded in the disciplines of the team members. Some authors argue that interdisciplinary research can only be done by experts in the disciplines (Billilign, 2013). That is, members of the interdisciplinary team must have a strong background in their chosen discipline and remain

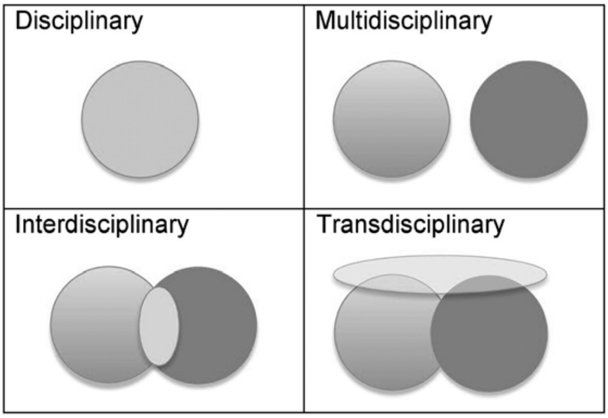


Figure 32.2 *Transdisciplinary model (Total Communication, 2019).*

active and current in their own field. Without depth of disciplinary knowledge and continuous updating, efforts to gain new perspectives and forge new understandings are compromised (Morley & Cashell, 2017).

In an effort to build interdisciplinary dialogue in public health, Collyer (2018) uses three metaphors to describe disciplinary training: a flashlight, a box, and a lens. As a *flashlight*, a particular disciplinary approach shines a light on specific aspects of the research problem at hand, leaving other aspects unlit (and unconsidered). This highlights the importance of having team members from different disciplines take time to talk about their approaches, share knowledge and perspectives, and uncover potential differences in their approaches. Disciplinary training can also be thought of as a *box* – a discipline's sense of "normal science." This allows a scientist to identify the "familiar in the unfamiliar" when considering new phenomena and provides them with the tools to make decisions regarding which research questions to study. The third metaphor is a *lens* through which researchers from a particular discipline see the scientific world, connecting meanings to concepts. Collyer recommends that interdisciplinary researchers take time to determine the key concepts that underlie the approaches to addressing a research question, recognizing that there may be differences in understanding of these concepts by discipline. It is also recommended that researchers be able to communicate about research methodology without relying on terminology specific to a discipline.

As the degree of integration across disciplines increases in a research collaboration, team members are likely to encounter differences in terminology, concepts, data analysis methods, theory, and even scientific assumptions (Urbanska et al., 2019). What is a researcher to do when what a collaborator from another discipline has written in a manuscript contradicts a theory from that person's own discipline? For example, there may be disagreement over the evidence required to demonstrate causality. Do the data show that a particular exposure led to a particular outcome? Given the same set of data, individuals from different disciplines may answer differently. In general, researchers are encouraged to note the difference and discuss it directly with their fellow researchers. If there is still a difference, team members may need to come to a place where they can agree to disagree. As recommended previously, research teams are encouraged to identify

Case Study A: A single discipline team transforming into a transdisciplinary team

A Quiet Firehouse (Carey et al., 2018)

Reducing Environmental Stimuli Among Professional On-Duty Firefighters

A nursing team of investigators lead by Carey and co-investigators revealed that, among on-duty firefighters, over half had elevated, average heart rates – a high-risk electrocardiographic marker for fatal cardiac events (Al-Zaiti & Carey, 2015; Carey et al., 2010). For a decade, the nursing team published studies capturing the burden of firefighting on the cardiovascular system. As a result, firefighters from around the world contacted the team, seeking professional input on their practices and, in one case, the fire department requested the team study the effects of fire alarms on firefighters in the firehouse. In this case, the nursing research team transformed into a transdisciplinary team that included non-scientist stakeholders – firefighters. In the USA, the fire service is designed as a paramilitary workforce; thus, it was important to include firefighters of all ranks including the commissioner, chief, captain, lieutenant, and officers because each rank would have independent input for the researchers.

Briefly, firehouse alarms are so loud that they cause a systemic response, similar to the flight-or-flight response. The purpose of the study was to reduce firehouse environmental stimuli, to improve sleep quality, and, thus, reduce cardiac burden. The firehouse intervention included restricting unnecessary fire alarms, reducing light levels, and regulating temperature in the bunkroom. Six weeks after implementing the interventions, measures revealed the average lux light level dropped from 0.75 to 0.19 lux, $p < 0.05$, and the presence of elevated blood pressure reduced from 86% to 15%, $p < 0.05$. The study results supported that reducing environmental stimuli in firehouses reduces blood pressure – a proxy for cardiovascular burden. On the basis of this pilot study, it was recommended that the practice of routinely activating unnecessary fire alarms in firehouse bunkrooms should be discouraged. The traditional nursing team developed into a genuine transdisciplinary team that published the results of the study; the fire captain and a public health practitioner were co-authors on the manuscript (Carey et al., 2018).

a process for handling potential points of difference, such as these, at the outset of a research project. Differences that come up in the collaborative process can then be handled according to the previously determined plan for handling conflicts.

One way to provide an example is through case studies; Case Study A comes from the disciplinary field of nursing.

Non-Traditional Research Collaborations

The benefits of non-traditional collaborations include bringing a broader awareness to research (Dick, 2017). The expression "teamwork makes dream work" comes to mind when collaboration is intentionally formed to generate diversity among team members, including across disciplines, experience, and geography. It is important to be open to non-traditional collaboration opportunities and to understand that challenges will need to be addressed early on in the collaboration; frequent meetings and open communication can reduce tensions among team members.

Community-Based Participatory Research: A Type of Fruitful Collaboration

Traditionally trained scientists may conduct thousands of experiments and publish hundreds of publications without actually thinking of their research participants. That is because scientists are trained that a participant is equivalent to an "*n*" or number; depending on the phenomena of interest and the subsequent effect size, there is a target *n* for each study. The goal is to enroll that number of participants to have the statistical power to test the research hypothesis. Theoretically, it does not matter if the participants are heart-failure patients, high-school students, or purple cows – they are just *n*. This strictness exercised by the scientists ensures objectivity and reduces bias in the research process (e.g., subject recruitment, interpretation of the data).

In contrast, community-based participatory research (CBPR) was developed based on the model of participatory action research (PAR) (Jull, 2017). PAR includes research participants, named as such because they participate in the research process as a research team member. PAR recognizes that participants possess knowledge and experience that are able to significantly contribute to the research process. For example, if an investigator is seeking to recruit undocumented migrant farm workers, it may be helpful to have an undocumented migrant farm worker as a research team member to advise how best to access this vulnerable workforce. PAR and CBPR is controversial because it democratizes the research process, so that research subjects work in collaboration with the scientists; this significantly changes the relationship between researchers and participants.

A CBPR project begins with the community; this includes a geographic community, a community of individuals with a common problem, or a community of individuals with a common goal. For example, an urban community may be interested in implementing healthy foods into their existing corner front stores. CBPR encourages collaboration of partners from any area of expertise that is seen as useful to the investigation as long as they are fully committed to a partnership of equals and producing outcomes usable to the community. True CBPR results in equitable partnerships by sharing power, resources, credit, results, and knowledge. This also involves a reciprocal appreciation of each partner's knowledge and skills, at each stage of the research project (e.g., research design, conducting research, interpreting the results, and strategies for implementation). Traditional research advances disciplinary knowledge while CBPR is an iterative process, incorporating research, reflection, and action in a cyclical process.

Case Study B further expands on Case Study A and provides a real-time example of CBPR.

Best Practice Models for Fruitful Collaboration

Education

There have been calls for training in interdisciplinary research from different fields (Gill et al., 2015; Khoury et al., 2013). Competencies for interdisciplinary health research fall into three domains: research conduct, communication, and interaction with others (Gebbie et al., 2008). For example, new researchers should be able to draw on theories and methods from multiple disciplines and integrate them into research approaches. They should also be able to collaborate with researchers from other disciplines on grant proposals and publish findings from interdisciplinary work outside of one’s discipline.

Research

There are numerous activities that constitute research productivity and include grantsmanship, research protocol development, data collection and analysis, manuscript development, scientific training, research dissemination, etc. When it comes to publishing, disagreements among members of interdisciplinary teams related to authorship are common, in part, because author order is typically determined by discipline-specific norms and conventions (Smith & Master, 2017). In some disciplines, authors are listed in order of degree of contribution (i.e., with the senior author listed last) while in other disciplines the authors are listed in alphabetical order. In addition to the order of authors being a point of contention, and representing a difference in convention between disciplines, who meets the criteria to be included as an author is a common point of contention.

Traditionally, to earn authorship, all authors must be involved in the drafting of the manuscript (International Committee of Medical Journal Editors, 2022). The requirement to contribute to the preparation of a manuscript may be difficult for many research teams because the manuscript reports the outcome of the research that involved important contributions from many team members. It is common to have a team member who takes the lead on drafting and preparing the manuscript for final approval, but without the contribution of others (e.g., data collectors) the study results would not be publishable. Importantly, many academics’ career success depends on a consistent publication record. In other words, research team members need to be recognized for substantial contributions, even when these contributions were not to the writing; thus, some have proposed recognizing their contributorship rather than authorship (Rennie et al., 1997). Indeed, journals generally require statements of contributorship that specify the roles of each of the authors.

Smith and Master provide a detailed review of differences in authorship and potential points of conflict in interdisciplinary collaborations; the result is a five-step best practice for determining contributorship and authorship order, intended for multi/interdisciplinary teams in academic health research (Smith & Master, 2017). These steps will look familiar, as they repeat themes previously mentioned for effective research collaboration. The first step involves delineating roles and responsibilities of the team members at the beginning of the research project as well as coming up with a process for resolving conflicts among team members. Second, still early in the project, team members should determine an order for authorship that is based on the extent of contribution (i.e., those contributing the most would be listed first). One thing the authors note is that it’s important to ensure that one discipline does not privilege one type of work over another (e.g., writing over data analysis).

Case Study B: Community-based participatory research – the pros and cons

A Quiet Firehouse (Carey et al., 2018)

Reducing Environmental Stimuli among Professional On-Duty Firefighters

Using the same example that was presented in Case Study A, the study, "A quiet firehouse" applied the principals of CBPR by including a fire captain on the research team. There were numerous strengths to including the fire captain. For example, when recruiting firefighters to participate in cardiovascular research, the captain advised the team what possible physiological measures would be amenable to collect. Instead of obtaining blood and transporting it to the lab, he suggested that the firefighters would prefer point-of-care testing. With immediate results and observation of the destruction of their blood sample, this ensured there was not deceptive research testing (e.g., drug testing). This highlights the importance of understanding the nuances of the group that an investigator is studying. As an outsider, a researcher may not understand what inhibits subjects from volunteering; thus, having the perspective of the participants improves participation rates. Increasing participation rates improves generalizability of the research findings because the sample collected from the potential population better represents the group. Thus, the research findings are closer to the truth.

As a nurse scientist (MGC), I recruited firefighters to participate in cardiovascular research. My approach was "field research" where I took my team and research equipment to the firehouse and conducted the study in the "field." This was effective because firefighters are obligated to stay at the firehouse while they wait for fire calls to be deployed. If the firefighter is not busy with work-related tasks, they may have time to participate in a research protocol. To protect the firefighters' health information, all studies were conducted anonymously; thus, when the first firefighter volunteered, they were given a participant number. My team would stay at the firehouse until all firefighters who wanted to participate in the study protocol had the chance to volunteer.

All of the data were analyzed in sequential or numerical order. When I completed studying the firefighter's results at the university, I returned to the firehouse to meet with each individual firefighter to share and explain the results. I began in the kitchen of the firehouse with all of the firefighters gathered and asked to meet with them individually in the firehouse bunkroom to ensure privacy; on average, I met with about five firefighters a week. In one case, I returned to the firehouse and was meeting with participant 28. An African American firefighter said, "Hey, Doc; why are you only giving the white guys their results?" I paused because I had not noticed the pattern. Then I replied, "Well Joe, I'm just returning the results in the same order that you guys volunteered to do my study, so it looks like the white guys volunteered first." Joe replied, "Hum, makes sense because black folk don't trust researchers, so I had to check you guys out for a while before I decided to volunteer."

With time, I eventually had a representative sample of the population of firefighters. My initial "field work" design to conduct the research in the firehouse was to optimize feasibility, but now it was apparent that it also optimized enrollment of minority subjects who needed the additional time to develop trust and confidence in my team. We essentially avoided the "helicopter" approach of quickly landing, grabbing the data and taking off. That just leaves any research subject feeling "used." Another challenge in recruiting minorities is that, more than 90% of the time, the face of the investigator is white and privileged. Correcting that imbalance will take substantial efforts, beginning in high school, to recruit students of color to pursue scientific careers. In the meantime, we provided a presentation on strategies for ethnic minority recruitment and retention in clinical research at the Annual Health Professions' Faculty Colloquium and our Annual Diversity Seminar Series; the presentation was well attended as researchers struggle with ethnic diversity of their volunteers (Box 32.1).

Case Study B: *(cont.)*

Box 32.1 Strategies to improve enrollment of underrepresented minority subjects

(1) The ethnic composition of the research team should reflect the population being recruited so consider including people on research team that are of community including churches, barber shop, etc.
(2) Understand the incentives of the population to participate: cash, international telephone call, dollar store gift cards
(3) Use CPBR strategies
(4) Include the principal investigator on enrollment strategies; have the principal investigator meet and thank subjects

While there were numerous advantages to CBPR, there were challenges when the study results were unflattering of the fire service. In one case, the results showed that nearly 70% of the firefighters had metabolic syndrome – large waist circumferences, high blood pressure, high cholesterol, and high glucose levels. Understandably, the fire captain did not want to publish data that revealed that the firefighters may not be "fit for duty." After extensive discussion within the team, it was agreed the data would be published and recommendations to assist the firefighters were provided to the union and fire service headquarters. Thus, with any approach, there are strengths and limitations – the art of balancing the two sides to make conclusions and recommendations based on the science is essential.

The team would also determine who will have contributed sufficiently to warrant authorship versus acknowledgement.

The third step is to have continuing dialogue about authorship, as the research process continues, accommodating changes as they arise (e.g., changes in team membership). The fourth step is to make a final determination on contributorship and authorship as the manuscript is being finalized for journal submission. The fifth step involves writing the statement on contributorship for inclusion in the manuscript. "The idea is that contributorship should reflect authorship and both should be declared, ideally in the manuscript, explaining and justifying authorship order" (Smith & Master, 2017, p. 259).

Another potential point of contention in interdisciplinary research is coming to consensus on where to publish the results of the research. As with authorship, this decision may be influenced by disciplinary norms and needs to be decided early in the research process. Team members will need to come to consensus on journals to target. Some may want to target a high-impact journal while others may be more interested in getting the research published in a journal from their own discipline or targeting a journal that can publish their research in a timely fashion. Some of this will be determined by the particular research question at hand. In interdisciplinary and transdisciplinary teams, the journal may be out of a researcher's usual comfort zone; it is possible that the team member will not get the same benefits they would, had they published within their discipline.

Service

One of the most rewarding and fruitful forms of scientific “service” is contributing and serving professional academic societies. For example, the Cardiovascular Council for Nursing at the American Heart Association is a council dedicated to cardiovascular nursing within the larger association dedicated to cardiovascular health. These national associations are well structured and supported by career staff who assist scientists to participate in national priorities. For example, the American Heart Association prepared a statement on psychological health, well-being, and the mind–heart–body connection (Levine et al., 2021). The statement encourages clinicians to treat the person rather than just the disease – to specifically provide more attention to psychological health and how that contributes to physical health and disease. Important services that also generate fruitful collaborations includes school- or university-level services that provides opportunities for co-workers to meet and to exchange research ideas that may be of interest to multi-participants. Most science is self-regulated, meaning it depends on the peer-review process. Thus, an important form of service is providing high-quality peer review of all forms of dissemination (e.g., journal articles, conference abstracts, grant applications) to ensure accurate and reliable information.

Tools of Collaborative Research

Collaborative work can take more time and requires good communication and project management (Nyström et al., 2018). Thus, numerous tools have been designed to reduce the hassle of collaboration while optimizing the productivity of the team.

Checklists

Gavens and colleagues have developed a “good-practice checklist” to support interdisciplinary public health research (Gavens et al., 2018). The checklist contains items that fall into five domains: blueprint, attitudes, staffing, interactions, and core science (BASIC). Briefly, interdisciplinary projects need to have an agreed-on project plan, a feasible scope and timeline, and flexibility to address unforeseen challenges (blueprint). There needs to be agreement among the team members regarding the importance of interdisciplinary work (attitudes). Staffing for the project needs to include elements of redundancy to address changes in team membership that may occur over time (staffing). In the staffing domain, they note the importance of assigning “interdisciplinary facilitators” who lead the task of synthesizing interdisciplinary findings. The checklist includes recommendations for interactions that facilitate interdisciplinary work – recommending “frequent, interactive, face-to-face meetings” and a “participative approach giving equal status to all disciplines” (interactions) (Gavens et al., 2018, p. 180). Finally, team members need to agree on core concepts and criteria for scientific evidence (core science).

Timelines and Publication Plans

Many scholarly activities have “due dates” associated with the activity, which include grants, school papers, invited papers, presentations, abstracts, book chapters, etc. With these types of activities, it is useful to create a timeline that starts with the deadline and works backwards; to keep everyone on track, list the required elements of the work, who is assigned, and interim deadlines for each step. Some scholarly activities do not have due dates (e.g., principal investigator-initiated data-based manuscripts). These manuscripts often include new data and are prepared throughout the year to contribute to the literature and to demonstrate a lab’s productivity. In these cases, it helps to create a due date to guide the authors in developing the work. Having a timeline can also help prevent projects from languishing when there isn’t a firm due date.

Creating a “publication plan” for the year helps manuscript development, review, and revisions stay on track. An example of a publication plan for building fruitful collaborations is provided in Table 32.1. It can be individualized, as needed; in this example, the columns include authors, title, journal, status, publication date. The publication plan helps track the progress of a manuscript and ensures a manuscript reaches publication; this takes time and perseverance. For example, the manuscript, “Hospital-based research internship for nurses: The value of academic librarians as co-faculty” (Carey et al., 2019) was rejected by two journals before it was finally accepted by *Journal of Nurses Professional Development*. Importantly, the manuscript was first submitted in May 2018, finally accepted for publication by February 2019, and was published by October 2019, representing an 18-month cycle; this is not unusual with dissemination in the healthcare sciences.

Collaborative Platforms

Research collaboration is growing exponentially, and teams are becoming ever more interdisciplinary, as researchers increasingly work in multidisciplinary and international groups to provide diversity in perspectives on research problems. Historically, collaboration took place in person at conferences, seminars, and campus visits. They often included socializing over meals and celebratory award banquets to highlight the annual accomplishments of members. However, given the recent online work trend, collaborative platforms depend on virtual platforms with demonstrated success. For example, the internet facilitated the first Virtual Chest Wall Injury Society summit (Sarani et al., 2021). The annual medical peer meeting survey results reported, among 275 registered participants, most participants (84%) felt the educational quality was excellent or good; however, most (75%) felt that in-person meetings were still better for education and networking and 87% preferred an in-person meeting in the future but would attend a virtual meeting again. Thus, collaborative platforms are virtual environments that provide research teams the opportunity to cooperate on their research, such as sharing research experience and ideas. Exercising national and international research collaboration is an expectation in the current virtual world.

Table 32.1 *Example of a publication plan for collaboration*

	Authors	Title	Journal	Status	Publication date
1	Carey, Qualls, Burgoyne	Patient perception of stressful events in the ICU following cardiac surgery	*American Journal of Critical Care*	Submitted May 2018 Accepted	Oct 2018
2	Carey, Trout, Qualls	Hospital-based research internship for nurses: The value of academic librarians as co-faculty	*Journal Medical Librarian Association* *Journal of Nursing Scholarship* *Journal of Nurses Professional Development*	Submitted May 2018 Rejected Submitted Aug 2018 Rejected Submitted Feb 2019 Accepted	Oct 2019
3	Carey, Nowzari, Finnell	A brief video intervention to teach firefighters the neurobiological basis of high-risk alcohol use	*Substance Abuse Journal of Psychiatric and Mental Health Nursing, J of Am Psych Assoc Journal of Studies in Alcohol and Drugs*	Submitted Aug 2013, Rejected Submitted May 2018, Rejected Submitted Aug 2018 Rejected Submitted Feb 2018 Accepted	Aug 2019
4	Carey & McMullen	The value of using an acuity score for neonatal nursing research	*Nursing Research* *Advances in Neonatal Care* *Journal of Nursing Scholarship* *Journal of Perinatal and Neonatal Nursing*	Submitted Sept 2017 Rejected Submitted May 2018 Rejected Submitted Jan 2019 Rejected Submitted May 2019 Accepted	Jan 2020
5	Fearrington, Qualls, Carey	Essential oils to reduce post-operative nausea and vomiting	*Journal of PeriAnesthesia Nursing*	Submitted Aug 2019 Accepted	Oct 2019

Research teams, particularly those from different institutions, may need to find ways to securely share data and, as a part of this process, will need to develop data-sharing agreements. There are online tools available that can facilitate collaboration, including sharing of research instruments and data. REDCap (Research Electronic Data Capture, Vanderbilt University) is a secure web-based platform for building surveys, collecting, and sharing data (Harris et al., 2009). Importantly, it provides regulatory compliance for securing data and provides vast support for a network of international researchers. The Open Science Framework, developed by the Center for Open Science, is a free, open-source project management tool that allows researchers to share a virtual workspace, files, and data (Foster & Deardorff, 2017). It supports add-ons for citation management and storage. All of the commonly used citation managers, including Endnote (Clarivate), Refworks (Ex Libris), Zotero (free and open source), and Mendeley (Elsevier), allow for sharing libraries between users (Perkel, 2020). With reference software, generating electronic libraries with thousands of references becomes automated and nearly error-free because there is no manual transferring of information.

Conclusion

Collaborative research is encouraged and comes with many benefits, including the development of professional relationships and, optimally, the enjoyment of working with a team. Interdisciplinary research, in particular, can be intellectually stimulating. There are key elements that increase the likelihood that a research collaboration will bear fruit. This includes trust, clear communication, common expectations, and an openness to learning from others. Best practices include being proactive, at the outset of the research collaboration, in planning how the team will work together, communicate, handle conflict, and share recognition.

Different types of research approaches – multidisciplinary, interdisciplinary, and transdisciplinary – are characterized by their degree of integration across disciplines. Transdisciplinary research, the most integrated of the three types, may involve researchers from different scientific disciplines, non-scientists, and the participants of the research (e.g., CBPR). Working on an interdisciplinary team comes with challenges but also provides the capacity to solve problems that are too big for one scientific discipline. If done successfully, building fruitful collaborations yields a project in which, in the words of Aristotle, “the whole is greater than the sum of its parts.”

References

Al-Zaiti, S. & Carey, M. (2015). The prevalence of clinical and electrocardiographic risk factors of cardiovascular death among on-duty professional firefighters. *Journal of Cardiovascular Nursing*, *30*(5).

Atkins, P. W. B. & Wood, R. E. (2002). Self- versus others' ratings as predictors of assessment center ratings: Validation evidence for 360-degree feedback programs. *Personnel Psychology*, *55*(4), 871–904. https://doi.org/10.1111/j.1744-6570.2002.tb00133.x

Bennett, L. M. & Gadlin, H. (2012). Collaboration and team science: From theory to practice. *Journal of Investigative Medicine*, *60*(5), 768–775. https://doi.org/10.2310/JIM.0b013e318250871d

Billilign, S. (2013). *The Need for Interdisciplinary Research and Education for Sustainable Human Development to Deal with Global Challenges*. North Carolina A&T State University.

Breckler, S. (2005). The importance of disciplines. Available at: www.apa.org/science/about/psa/2005/10/ed-column.

Breland, A., Balster, R. L., Cobb, C., et al. (2019). Answering questions about electronic cigarettes using a multidisciplinary model. *American Psychologist*, *74*(3), 368–379. https://doi.org/10.1037/amp0000426

Bridges, D. R., Davidson, R. A., Odegard, P. S., Maki, I. V., & Tomkowiak, J. (2011). Interprofessional collaboration: Three best practice models of interprofessional education. *Medical Education Online*, *16*. https://doi.org/10.3402/meo.v16i0.6035

Carey, M., Al-Zaiti, S., Liao, L., Butler, R., & Martin, H. (2010). Characteristics of the standard 12-lead holter ECG in professional firefighters. *Computing in Cardiology*, *37*, 685−688.

Carey, M., Baldzizhar, A., Miterko, C., et al. (2018). A quiet firehouse: Reducing environmental stimuli among professional on-duty firefighters. *Journal of Environmental Medicine 60*(2), 186–190.

Carey, M. G., Regehr, C., Wagner, S. L., et al. (2021). The prevalence of PTSD, major depression and anxiety symptoms among high-risk public transportation workers. *International Archives of Occupational and Environmental Health*, *94*, 867–875. https://doi.org/10.1007/s00420-020-01631-5

Carey, M. G., Trout, D. R., & Qualls, B. W. (2019). Hospital-based research internship for nurses: The value of academic librarians as cofaculty. *J Nurses Prof Dev*, *35*(6), 344–350. https://doi.org/10.1097/nnd.0000000000000585

Choi, B. C. & Pak, A. W. (2006). Multidisciplinarity, interdisciplinarity and transdisciplinarity in health research, services, education and policy: 1. Definitions, objectives, and evidence of effectiveness. *Clinical and Investigative Medicine. Medecine Clinique et Experimentale*, *29*(6), 351–364.

Collyer, T. A. (2018). Three metaphors to aid interdisciplinary dialogue in public health. *American Journal of Public Health*, *108*(11), 1483–1486. https://doi.org/10.2105/ajph.2018.304681

Davis, J. (1995). Interdisciplinary courses and team teaching: New arrangements for learning *Assessment and Evaluation in Higher Education*, *22*(3), 348–350.

Dick, D. M. (2017). Rethinking the way we do research: The benefits of community-engaged, citizen science approaches and nontraditional collaborators. *Alcoholism, Clinical and Experimental Research*, *41*(11), 1849–1856. https://doi.org/10.1111/acer.13492

Foster, E. D. & Deardorff, A. (2017). Open Science Framework (OSF) [Product Review]. *Journal of the Medical Library Association*, *105*(2), 203–206.

Garcia-Dia, M. J. (2021). Nursing informatics: An evolving specialty. *Nursing Management*, *52*(5), 56. https://doi.org/10.1097/01.NUMA.0000743444.08164.b4

Gavens, L., Holmes, J., Bühringer, G., McLeod, J., et al. (2018). Interdisciplinary working in public health research: A proposed good practice checklist. *Journal of Public Health (Oxf)*, *40*(1), 175–182. https://doi.org/10.1093/pubmed/fdx027

Gebbie, K. M., Meier, B. M., Bakken, S., et al. (2008). Training for interdisciplinary health research: Defining the required competencies. *Journal of Allied Health*, *37*(2), 65–70.

Gill, S. V., Vessali, M., Pratt, J. A., et al. (2015). The importance of interdisciplinary research training and community dissemination. *Clinical and Translational Science*, *8*(5), 611–614. https://doi.org/10.1111/cts.12330

Goldsberry, J. W. (2018). Advanced practice nurses leading the way: Interprofessional collaboration. *Nurse Education Today*, *65*, 1–3. https://doi.org/10.1016/j.nedt.2018.02.024

Harris, P. A., Taylor, R., Thielke, R., et al. (2009). Research electronic data capture (REDCap): A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, *42*(2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010

International Committee of Medical Journal Editors (2022). Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals (accessed September 2022).

Jull, J., Giles, A., & Graham, I. D. (2017). Community-based participatory research and integrated knowledge translation: Advancing the co-creation of knowledge. *Implementation Science*, *12*(1), 150. https://doi.org/10.1186/s13012-017-0696-3.

Khoury, M. J., Lam, T. K., Ioannidis, J. P., et al. (2013). Transforming epidemiology for 21st century medicine and public health. *Cancer Epidemiology, Biomarkers and Prevention*, *22*(4), 508–516. https://doi.org/10.1158/1055-9965.Epi-13-0146

Klein, J. T. (1990). *Interdisciplinarity: History, Theory, and Practice*. Wayne State University Press. https://psycnet.apa.org/record/1990-97814-000

Levine, G. N., Cohen, B. E., Commodore-Mensah, Y., et al. (2021). Psychological health, well-being, and the mind–heart–body connection: A scientific statement from the American Heart Association. *143*, e763–e783. https://doi.org/doi:10.1161/CIR.0000000000000947

Lindner, M. D., Torralba, K. D., & Khan, N. A. (2018). Scientific productivity: An exploratory study of metrics and incentives. *PloS One*, *13*(4), e0195321–e0195321. https://doi.org/10.1371/journal.pone.0195321

Moradian, N., Ochs, H. D., Sedikies, C., et al. (2020). The urgent need for integrated science to fight COVID-19 pandemic and beyond. *Journal of Translational Medicine*, *18* (1), 205. https://doi.org/10.1186/s12967-020-02364-2

Morley, L. & Cashell, A. (2017). Collaboration in health care. *The Journal of Medical Imaging and Radiation Sciences*, *48*(2), 207–216. https://doi.org/10.1016/j.jmir.2017.02.071

National Academy of Sciences, National Academy of Engineering, & Institute of Medicine (2005). *Facilitating Interdisciplinary Research*. The National Academies Press. https://doi.org/doi:10.17226/11153

National Research Council (2014). *Convergence: Facilitating Transdisciplinary Integration of Life Sciences, Physical Sciences, Engineering, and Beyond*. The National Academies Press. https://doi.org/doi:10.17226/18722

Newell, W. & Klein, J. (1996). Interdiscplinary studies into the 21st century *Journal of General Education*, *45*(2), 152–169.

Nyström, M. E., Karltun, J., Keller, C., & Andersson Gäre, B. (2018). Collaborative and partnership research for improvement of health and social services: Researcher's experiences from 20 projects. *Health Research Policy and Systems*, *16*(1), 46. https://doi.org/10.1186/s12961-018-0322-0

Peek, L. & Guikema, S. (2021). Interdisciplinary theory, methods, and approaches for hazards and disaster research: An introduction to the special issue. *Risk Analysis*, *41*(7), 1047–1058. https://doi.org/10.1111/risa.13777

Perkel, J. M. (2020). Streamline your writing – and collaborations – with these reference managers. *Nature*, *585*(7823), 149–150. https://doi.org/10.1038/d41586-020-02491-2

Pescosolido, B. A., Perry, B. L., Long, J. S., et al. (2008). Under the influence of genetics: How transdisciplinarity leads us to rethink social pathways to illness. *AJS: American Journal of Sociology*, *114*, S171–201. https://doi.org/10.1086/592209

Reeves, S., Pelone, F., Harrison, R., Goldman, J., & Zwarenstein, M. (2017). Interprofessional collaboration to improve professional practice and healthcare outcomes. *Cochrane Database Syst Rev*, *6*(6), CD000072. https://doi.org/10.1002/14651858.CD000072.pub3

Rennie, D., Yank, V., & Emanuel, L. (1997). When authorship fails. A proposal to make contributors accountable. *JAMA*, *278*(7), 579–585. https://doi.org/10.1001/jama.278.7.579

Repko, A. (2008). *Interdisciplinary Research: Process and Theory*. SAGE Publications.

Sarani, B., Shiroff, A., Pieracci, F. M., et al. (2021). Use of the Internet to facilitate an annual scientific meeting: A Report of the first Virtual Chest Wall Injury Society summit. *Journal of Surgical Education*, *78*(3), 889–895. https://doi.org/10.1016/j.jsurg.2020.09.004

Smith, E. & Master, Z. (2017). Best practice to order authors in multi/interdisciplinary health sciences research publications. *Accountability in Research*, *24*(4), 243–267. https://doi.org/10.1080/08989621.2017.1287567

Stokols, D., Hall, K., Taylor, B. K., & Moser, R. P. (2008). The science of team science. *American Journal of Preventive Medicine*, *35*, S78–S89.

Szostak, R. (2002). How to do interdisciplinarity: Integrating the debate. *Issues in Integrative Studies*, *20*, 103–122.

Total Communication (2019). Transdisciplinary approach: What does it mean? Available at: www.totalcommunication.com.sg/post/transdisciplinary-approach-what-does-it-mean (accessed February 16, 2021).

Urbanska, K., Huet, S., & Guimond, S. (2019). Does increased interdisciplinary contact among hard and social scientists help or hinder interdisciplinary research? *PloS One*, *14*(9), e0221907.

VanNoorden, C. (2014). Interdisciplinary research by the numbers. In B. Cronin & C. Sugimoto (eds.), *Beyond Bibliometrics* (p. 480). MIT Press.

33 Performing a Good Peer Review

Klaus Fiedler and Christian Unkelbach

Abstract

Peer review supports decisions related to publications, grant proposals, awards, or personnel selection. Independent of the specific occasion, we propose validity as a chief evaluation criterion for reviews. While applicable to all occasions, the principles of validity-oriented quality control are particularly suited for journal reviews. Beyond evaluating validity and the scientific potential of a given piece of research, we address how peer reviewing serves important functions and is accountable for the growth of science at a more superordinate level. We also provide guidelines and concrete recommendations for how a good peer review may serve these functions. Good peer review, thereby, fosters both the advancement of scientific research and the quality, precision, and sincerity of the scientific literature. The end of the chapter is devoted to a core set of good reviewer practices, conceived as an essential feature of academic culture.

Keywords: Internal Validity; External Validity; Diagnostic Design; Theoretical Priors; Advancement of Science; Writing Style; Good Researcher Practices

Introduction

Peer reviews are solicited for many different purposes in academia – from graduate admission decisions to grant proposals and allocation of scientific awards. The present chapter focuses on the prototype of reviewing for scientific journals. Journal editors ask experts in a given field for their advice in evaluating an article submitted for publication (see Chapter 34 in this volume). Peer reviews play a central role in this publication process because the expert reviewers' feedback often determines whether a manuscript is published or not. Thus, the review process determines which subset of documented research will be accessible to the scientific community. As the proportion of published articles is often as low as 20% or 10%, or even less than 10%, of all submissions, the review process has a crucial impact on the unfolding of a research discipline.

The journal review process has two main functions – advising editors and authors. It guides editorial evaluations and decisions, and it helps authors to shape and sharpen their contribution. Regarding both functions, the value of good peer reviewing cannot be overestimated. Peer reviewing is crucial for quality control in science, and it serves a major fertilization function. The beauty of some of the most compelling publications reflects, to a considerable degree, the wisdom and advice of anonymous reviewers. In the best case, a mixture of prosocial, advisory, competitive, and even self-presentational reviews shape a submitted manuscript into a masterful publication. In the worst case, a stubborn and narrow-minded review process can truncate the maturation process and prevent an article from unfolding its fascination.

In the following, we summarize what we consider essential conditions for realizing the former and avoiding the latter case. We first address some basic misconceptions about the function of peer review. Then, we outline different aspects of validity a reviewer should have in mind when doing a review. Having established these guiding principles, we address more concrete points for good reviewing and provide guidelines on how reviewers may fulfill their essential role as controllers and arbiters – but also as supporters and promoters of scientific advancement.

Basic Misconceptions of Peer Reviewing

Let us first try to get rid of a few basic misconceptions concerning peer reviewing. First, the peer-review process is often seen as a "gatekeeping" function. Accordingly, the reviewers' most prominent task is to keep misleading, erroneous, or blatantly false research from passing the publication gate, making a high rejection rate the chief quality criterion of a leading, highly selective journal. However, interpreting peer reviewing as a high-entry-threshold evaluation system may be counterproductive. A serious misunderstanding is the failure to note that false negatives (i.e., rejecting a good candidate manuscript) can be more expensive than false positives (i.e., accepting a poor candidate manuscript). Granting that the base-rate of truly excellent pieces of research, which entail ground-breaking innovations, is probably much lower than the base-rate of modest piecemeal research, too conservative a publication threshold is dysfunctional for the growth of science.

A cost–benefit analysis indicates that an optimal publication threshold must be more liberal, considering that outstanding research is rare, and not publishing the few outstanding ideas is costlier than erroneously publishing some weak findings. In other words, if the likelihood ratio of *p*(outstanding)/*p*(weak) is low and the utility ratio [Benefit(hits) + Cost(false negative)]/[Benefit(correct rejection) + Cost(false positive)] is high, it is important not to miss those rare and precious exemplars of outstanding research (see Swets et al., 2000). Framed in terms of scientific progress, by definition, only preventing poor research from being published cannot advance science. However, *good reviewing should be in the service of advancing science*. The resulting trade-off between both tendencies – high rejection rates and detrimental costs of false negatives – creates a heavy burden for good peer reviewing. Science cannot afford missing the most precious exemplars. Therefore, good peer reviewing must not be one-sidedly restrictive. Its ability to diagnose and meliorate the best ideas, and to avoid "misses," is at least as important as its sensitivity to detecting and rejecting mediocre or misleading examples (i.e., correct rejections).

Another misconception concerns the naïve ontological distinction of truth and falsehood. Publishing a research finding in a scientific journal does not bestow "truth" to a hypothesis or a manuscript. This is a common misunderstanding of the reviewers' and the editors' roles. Rather, reviewers and editors decide if a given manuscript is innovative, original, inspiring, theoretically elucidating, practically useful, or interesting for the readership of their journal or for the scientific community in general. Many true hypotheses may be trivial, well established, or unlikely to augment scientific knowledge. Conversely, a preliminary hypothesis that can be foreseen to be wrong or providing a seriously simplifying model of an ill-understood phenomenon can be inspiring and fascinating – affording a groundbreaking publication.

To be sure, a review must be sensitive to the latest "truth" – conceived as the state of the arts or the most recent approximation or update of an epistemic domain – as manifested, for example, in a pertinent review article. It entails scrutinizing whether the documented research follows the standards and good practices of a discipline, whether conclusions adhere to logic of science and methodology, and whether the strength of the evidence matches the strengths of the claims. If a manuscript violates good practices, draws illogical conclusions, or makes strong claims with little evidence, then it should not pass a review process. In this regard, peer review serves a quality-control function.

However, beyond quality assurance, the peer review process cannot and should not pass an ultimate answer concerning the truth or the falsity of a research hypothesis or a presented theory. In fact, reviewers would be hard-pressed to discriminate a false-positive from a true-positive finding without conducting or awaiting additional research (for a related discussion, see Edlund et al., 2022). This is the cumulative task for the scientific community – which may replicate, expand upon, or employ a given finding, hypothesis, or theory – and, thereby, substantiate or refute it. Oftentimes, the scientific community will not fulfill this function for each and every published article, but the scientific evolution will, at best, take up and develop a few open research questions, in a highly selective process. Yet, the basic function of a published manuscript, which can be supported through peer reviewing, is to invite the community to address and elaborate on a finding, a hypothesis, or a theory. Beyond this function, publishing a research finding does not make it "true."

Another common misunderstanding concerns the reliability or consistency of two, three, or even four reviews of the same article (Marsh & Ball, 1989). Because different reviews may complement each other, reflecting a division of labor among experts serving as consultants for different aspects (methods, theorizing, literature review), there is no logical need for all reviewers to arrive at the same evaluation. A memorable lesson from the advice-taking literature is that the quality of an advice taker's judgment increases when advice givers work independently, relying on non-overlapping sources (Yaniv et al., 2009). Thus, mutually complementary advice givers provide more valuable information than fully redundant consultants. However, importantly, relying on two or more reviewers who judge the same work from different perspectives calls for a sovereign referee – the editor – to integrate the differential perspectives into a coherent judgment (see Chapter 34 in this volume).

Clarifying Notes and Assumptions

Having cleared up common misunderstandings, let us be explicit about some assumptions to which we are committed throughout this chapter. First, it goes without saying that our positions concerning good reviewing reflect, to a considerable degree, two authors' opinions. Although the authors can claim to be quite experienced – as authors, reviewers, and editors – there is room for different opinions. Yet, we believe our opinions are substantiated by good arguments.

A clarifying remark is in order concerning different types of reviews embedded in academic decisions, in which different outcomes and consequences are at stake. If the target of a review is a submitted journal article, the evaluation threshold can vary; it may be a leading international flagship journal with a rejection rate above 90% or a specialized journal that is open for original publications of all kinds. The journal program also matters; some journals are confined to empirical research while others are open to reviews, theoretical comments, and adversarial discussions. If it is an invited book chapter, the reviewer's job is hardly to recommend an all-or-none decision; s/he should instead provide suggestions for how to improve the chapter's readability or fit to the overarching book project. If the target is a grant proposal, the question is whether financial investment is warranted and justified. If a review refers to a promotion, tenured position, professorial position, or another personnel selection problem, the purpose of the review is a fair comparison of personal competencies.

However, despite these different purposes, good reviews in such diverse areas resemble each other in one central way. Good reviews are concerned with the most essential criterion of scientific quality – validity. To be sure, many reviews also include comments on superficial aspects of text format, orthography, proper citation, readability and didactic quality, or technical aspects of statistical analysis. However, what authors, editors, and other reviewers have in mind when they praise a review as really good, constructive, and fair is typically how a review evaluates the validity of the documented research – the validity of the research depicted in a grant proposal or conducted by a scientist nominated for an award or an academic position. It is because of this validity focus, as a common denominator of all reviewing, that the final evaluation and recommendation will be very similar, regardless of the specific purpose of an invited review.

Validity-Oriented Peer Reviewing

The crucial distinctive feature of validity, distinguished from other, more formal and superficial evaluation criteria (e.g., readability, text style, compliance, or formal precision), is that validity taps the most relevant level for evaluating a research idea. Whereas other, less essential criteria (e.g., compliance with mainstream methods of analysis, linguistic style, formal notation, or proper citation of the extant literature) depend on preferences, conformity, power, and tolerance with pluralistic science norms, validity is at the heart of a piece of research. Therefore, assessing validity is the essence of fair evaluation, independent of reviewers' theoretical or methodological preferences. Validity problems cannot be resolved through rhetorical text reframing, re-labeling of a phenomenon, or politeness and obedience with editors or critical reviewers. Rather, validity provides the common ground for the scientific evaluation of a research idea's core, as opposed to its peripheral or superficial concomitants. Thus, with reference to Immanuel Kant's classical writings, one might say validity lies in the interface of the critique of pure reason and the critique of practical reason, where practical reasoning meets incontestable norms of fairness.

In the following, we provide a brief overview of what we consider relevant validity aspects that apply to most reviews in the social and behavioral sciences (see also Table 33.1).

Validity of Research Design

Our framing of validity – between pure and practical reasons – may sound lofty and a bit idealistic because even validity issues can be a matter of disagreement or debate; the advantage of validity-oriented reviewing is that even divergent standpoints have to be articulated in terms and argument structures that are more precise than in most other areas of scientific discourse. Thus, when peer reviewing revolves around, or provides suggestions to improve on, the validity issues summarized in Table 33.1, there is common ground that reviewers, editors, and authors cannot ignore or evade. They are jointly obliged to validity norms that are deeply rooted in the logic of science. Scientists are obliged to contemplate and clearly articulate their arguments for or against the notion of proper manipulation checks, sampling biases, measurement error, demand effects, or any of the other validity issues listed in Table 33.1. Nobody can simply deny or discard these issues as unfair or peculiar to arbitrary positions or belonging to certain camps or scientific groups with vested interests. Moreover, nobody can deny the relevance and pertinence of these issues in a quality-oriented evaluation process. In all these regards, validity is central to what we all expect of a good review and what gives meaning and justification to the review process.

Thus, a good reviewer's expertise, and the training experience of all players in the game, should first include Campbell's (1957) seminal work on internal and external validity. The internal validity of an experimental finding refers to all factors that speak to the crucial question of whether an observed change in a dependent variable Y (ΔY) was actually induced by an experimentally manipulated change in an independent variable X (ΔX). For instance, is a change in creativity actually due to a manipulated increase in positive mood, rather than to uncontrolled differences between experimental conditions?

Internal Validity. This basic validity issue involves such considerations as: (a) the comparability of experimental conditions in a randomized design; (b) the elimination or balancing of extraneous factors and uncontrolled influences other than X; (c) the proper understanding of the experimental instructions; (d) minimization of measurement error; and (e) the standardization of the experimental setting. Thus, to keep

Table 33.1 *Overview of recommended validity criteria that a good peer review should focus on*

Validity issue	Reason why it is important	Damage if neglected
Internal validity	Arguments related to the crucial question of whether change in a dependent variable is actually due to an experimentally induced independent variable change	Fundamental misinterpretation of an experimental result
External validity	Arguments related to the generality of findings across participants, stimuli, and task conditions	Unwarranted generalization claim
Convergent validity	Consistent results obtained with different methods used to test a hypothesis	Results peculiar to selective method or materials
Divergent validity	Specificity of findings that support a focal hypothesis more than rival hypotheses	Misattribution of common findings to focal hypothesis
Ecological validity	Diagnosticity of a set of observed or manipulated cues for a distinct hypothesis	Misinterpretation of results observed with distinct cues
Manipulation check	In an empirical test of the implication *if p, then q*, the premise *p* was established	Unwarranted rejection of an irrelevant hypothesis
Confound	Confusion of a focal variable with a similar or correlated variable	Confusion of variables used to denote an effect
Mediation	Explanation of a causal influence, $X \rightarrow Y$, in terms of an intermediate variable Z, $X \rightarrow Z \rightarrow Y$	Failure to detect the underlying causal chain
Demand effect	Change in the dependent variable induced by subtle demand cues conveyed in the instructions or stimulus materials	Misattribution of experimental results to independent variable

within the example, the internal validity of an experiment testing the impact of mood on creativity depends on the extent to which: (a) a positive and neutral mood condition are equivalent in all other respects; (b) mood is not confounded with any other relevant variable (such as achievement motivation); (c) a creativity test is correctly understood; (d) a proper measure of creativity is used; and (e) the task setting is kept constant across all participants and creativity tasks.

External Validity. Going beyond internal validity, external validity refers to all factors affecting the generalizability of internally valid findings across participants, stimuli, time and occasions, task settings, etc. Does the demonstration of enhanced creativity under positive mood generalize across participant groups (e.g., age, education, culture), creativity tasks, mood manipulations, task settings, etc.? The ideal case of an externally valid arrangement has been called a *representative design* (Brunswik, 1955) – treating participants, stimuli, occasions, and task conditions as random factors and making a design representative of naturally occurring correlations between all experimental factors.

Sources of Invalidity. It should be obvious that hardly any existing research covered in a single manuscript can live up to the criterion of external validity (Mook, 1983), whereas internal validity is applicable and central to virtually every piece of research. The failure to include a proper control condition or the failure to control for selective group assignment in a non-randomized design are mistakes that can be hardly corrected for. A proper manipulation check (Fiedler et al., 2021) is particularly important for internal validity; an empirical finding depends on whether a manipulation was actually effective in manipulating ΔX and not inadvertently manipulating other factors (ΔA, ΔB, ΔC, . . . etc.) – suggesting alternative explanations of the ΔY effect in terms of other causes than ΔX. A recent meta-analysis of manipulation checks in a full year of publications in the *Journal of Personality and Social Psychology: Attitudes and Social Cognition* (Fiedler et al., 2021) suggests that proper manipulation checks, especially in internet-based computer experiments, continue to be the exception rather than the rule, even in leading journals.

Another topic of an informed, open-minded validity analysis in a convincing review is alternative accounts related to demand effects – experimental instructions or cues of the task setting that tell participants (between the lines or even blatantly) what a good participant is supposed to do (see Chapter 11 in this volume). These and other entries in Table 33.1 exemplify ways in which an informed discussion of mundane validity issues – relying on Campbell's (1957) over 60-years old writings – can render a good review excellent and disarmingly convincing.

We invite readers to contemplate on Table 33.1 and add further useful entries, providing more examples of validity criteria that scientists must beware of. This may include the logical distinction of existence proofs (e.g., it is just possible that good mood may under auspicious conditions foster creativity) and universal proofs (e.g., positive mood may generally improve creativity). It may also be the difference of exploratory versus confirmatory research (e.g., the degree to which the hypothesized impact of mood on creativity can be derived on theoretical grounds).

Validity in Theorizing and Scientific Reasoning

So far, we have considered validity related to proper research designs. However, equally important for a good review is the insight that validity follows from proper theoretical reasoning and the logic of scientific rules. To illustrate the importance of theoretical constraints, consider another example – the socio-economic hypothesis that wealthy people (X) become powerful people (Y). Wealthier people should be more powerful compared to less wealthy people (i.e., $X \rightarrow Y$). Logically, this hypothesis implies that powerless people should be not wealthy (i.e., $\neg Y \rightarrow \neg X$; with $\neg Y$ and $\neg X$ denoting the negation of Y and X). However, it is important to understand that the rules of propositional logic do not constrain the other two possible conditional relations between X and Y. Less wealthy people (i.e., $\neg X$) may or may not be powerful, and powerful people (i.e., Y) may or may not be wealthy. For valid tests of these logically unconstrained relations, one would have to introduce the auxiliary assumption that being wealthy is the only causal condition that renders people powerful (i.e., $X \leftrightarrow Y$). A good peer review can be extremely helpful just by translating a narrative theory or hypothesis, stated in ordinary language, into a more precise propositional format.

Hempel–Oppenheim Scheme. In a similar vein, reframing a naïve theory within the deductive-nomological model of the so-called Hempel–Oppenheim scheme can be very helpful. Hempel and Oppenheim (1948) decomposed scientific explanations into three constituents. Given a theoretical law (e.g., wealthy people are powerful), and together with the assumption that a special case can be subsumed under the premise of that law (e.g., bankers are wealthy), the law applies to the special case. A review may greatly benefit from embedding the to-be-reviewed research within such propositional reasoning rules.

Quine–Duhem Problem. In the social and behavioral sciences, in particular, researchers must translate their theoretical constructs (e.g., wealth and power) into measurable variables. In the present case, disposable income could be used as a measure of wealth and executive functions as a measure of power. As a consequence, any empirical test of the relation ($X \rightarrow Y$) is not a pure test of the postulated theoretical relation but reflects, to an unknown degree, the specific means of operationalizing X and Y in terms of income and executive functions. This sort of indeterminacy is known as the Quine–Duhem problem (Duhem, 1954; Earp & Trafimow, 2015; Quine, 1980). The hypotheses laid out in a manuscript are often derived from previous research in which the same theoretical variables have been measured or operationalized in diverse and sometimes arbitrary ways. Good peer review should reveal the Quine–Duhem problem and not suppose naively that constant variable names must refer to identical underlying variables.

Validity at the Data Level

A very concrete validity level may be called statistical validity. To the extent that statistical procedures are becoming more and more complex and normative statistics more equivocal, the validity of statistical inference may constitute a challenge for good peer reviewing. Although it may be possible and advisable to involve statistical experts in the review process, a good peer review may be one that helps scientists evade such loss of control over statistical analyses of data.

While we cannot address the full wealth of potential pitfalls, we want to briefly address a prominent example frequently observed in the literature: absence of evidence is not the same as evidence of absence. If an unwanted influence has no detectable influence on the data (e.g., for our naïve theory example, one might aim to show that the wealth effect is independent of gender), one may not conclude that there is no influence. Logically, one cannot prove the non-existence of an influence (or any construct or relation, for that matter). With classic statistics, given such null effects, one must confine oneself to stating that, under the conditions of the given study, there is no statistically detectable influence. However, with the more recent usage of Bayes' statistics, it has also become possible to quantify the evidence for the absence of an influence (Wagenmakers et al., 2018).

Summary

The list above is far from exhaustive, and there are several other potential threats to the validity of research presented in each manuscript. We believe this list represents a primer for points of validity against which any manuscript may be evaluated. In addition, one may argue that some validity violations are defensible (e.g., lack of external validity in purely experimental research; Mook, 1983) while others are not (e.g., lack of internal validity for causal inferences). However, it is within each reviewers' individual judgment to delineate the consequences of a found validity violation for a given manuscript's publishability.

Beyond Validity: Other Important Review Aspects

Assessing validity is central to our analysis of what constitutes good peer review. However, we want to provide further guidance beyond the mere assessment of validity as a core mission of the review process.

A Focus on the Positive

Reviews are often frustrating for authors due to their focus on manuscripts' negative aspects. There are at least three reasons for this prevailing negativity. First, there are fewer ways in which manuscripts may excel compared to the many ways manuscripts may fail (see Alves et al., 2017; Unkelbach et al., 2019, 2020). Second, negative evaluations may reflect the evaluator's self-presentation strategies and his or her attempt to appear sharp and competent, committed to the highest standards of scientific excellence (Amabile, 1983). Third, most of the presented research may indeed be flawed (Sturgeon, 1957).

Independent of the reasons, a good review should try to overcome an exclusively negative focus – it should focus as much as possible on the positive. This shift may be achieved by pursuing and testing two competing hypotheses: "The presented research should not be published" versus "The presented research should be published." The first hypothesis is almost ingrained into the reviewer mission; here, we have also focused so far on potential support for this hypothesis in the form of validity threats. The second hypothesis necessitates a focus on a given manuscript's manifest or latent strengths. In other words, a good review should provide an editor with reasons to publish the manuscript. Ideally, these reasons go beyond a generic "the manuscript is well written" and "the methods are solid" assessment. Useful questions that help to uncover a manuscript's positive aspects are: Is the presented idea original? Is the employed method original? Are there innovative aspects? Does the research present new findings or replicate old ones? On the most general level, if the reviewer was a reader of a given journal, are there positive reasons why one should read it? Such questions lead to a mindset that brings a manuscript's strengths and positive potential into focus.

In addition, such mindsets foster a promotion focus (Higgins, 1997). As we have already delineated above, the most costly error in the reviewing process is due to misses – overlooking precious cases of good research that are not published – rather than to false alarms – bad research that might get published. Good reviewing should be marked by a promotion focus that fosters good research rather than a prevention focus that avoids bad research (Rosenthal & Rosnow, 1984).

A Focus on the Positive Does Not Imply Leniency. Although a good review should highlight the positive, it should not be lenient or uncritical. On the contrary, a good review is also obliged to honesty and transparency rather than friendliness and politeness norms. Fairness is not Pollyanna. Writing benevolent reviews and unrealistically positive recommendations for every research paper may appear philanthropic, but it is unfair vis-à-vis those who may present superior research (e.g., related to the validity points we delineated above). In social psychological terms, fair evaluation is a matter of equity (i.e., everybody gets a fair share) not equality (i.e., everybody gets the same). Indeed, it is common that authors are grateful for an elucidating review process that ended with a rejection decision rather than a happy, but potentially flawed, outcome.

Trade-off Between Two Maxims. Thus, review writing entails a trade-off between two maxims: (1) doing everything to discover the most noteworthy, hidden, or visible, value inherent in each piece of work but (2) being also highly sensitive to the discriminant value of different manuscripts. A good review succeeds in solving this trade-off by (a) doing everything to work out the positive potential of a piece of research, and (b) at the same time explaining and communicating the merits or deficits of research in an upfront and transparent way – according to the state-of-the-art standards in a field.

Informing Editorial Decisions

A review provides a clear recommendation for editors on what to do with a given manuscript. What is perhaps more important, a good review also provides clear reasons for this recommendation. The argumentation – the delineation of a manuscript's strengths and weaknesses – should be comprehensible and transparent. In this sense, a good review is highly similar to a good research paper.

The same way empirical research papers follow a structure of introduction, methods, and discussion, to facilitate communication, structuring a review facilitates communication between reviewers, authors, and editors. A standard format involves the typical essay form of an argument; this includes presenting a brief summary of the reviewed research followed by arguments of why this research should be published (i.e., the thesis), arguments why this research should not be published (i.e., the antithesis), and a final recommendation – based on the weighting of the previous arguments (i.e., the synthesis). In addition, structuring tools (e.g., numbering of arguments, using paragraphs, and using headlines) facilitate the editor's use of a given review.

To be sure, there are no fixed templates; just as experienced authors are not chained to the straightjacket of a standardized manuscript format, experienced reviewers may feel shackled by such formalities. However, for most people, most of the time, a shared canonical format facilitates comprehension and communication. Any deviation from the canonical format, if necessary, can be explained in a confidential comment to the editor – that should, however, remain an exception.

A Note on Length

The same way a review is similar to a research paper with regard to a common structure, similar rules of length apply. A review should be as long as necessary but as short as possible. An asymmetry for length often arises on the negative side, there is no need to present the counter-evidence when the reasons for rejecting a manuscript are clear and straightforward. For example, when a research report for an experimental journal violates the primary rules of internal validity with its procedures, it does not matter whether the manuscript is beautifully written, the statistical procedures are state of the art, and the samples are large and representative. If principles of internal validity or logic are violated, it also does not matter if the reference list is incomplete, and the writing structure is lacking.

On the other hand, if a manuscript is a candidate for publication – there is substantial evidence for the thesis that it should be published – a good review should and must address all the potential shortcomings to allow for the best version of the manuscript to be published. In passing, we note that this positive–negative asymmetry might be a reason for the prevalence of negative reviews; they are typically shorter and less time-consuming, highlighting only a few negative points. Positive reviews need to address all suboptimal aspects to help publish as good a manuscript as possible.

Likert-Scale Ratings of Manuscripts

Many journal submission systems ask reviewers to rate manuscripts on several dimensions (e.g., originality, methodological soundness, information-to-length ratio). Although such attempts of quantifications are popular, we want to advice against such routines, as the scales only provide an illusion of objectivity and validity. Such ratings are a matter of framing, labeling, comparison standard, and calibration – they are most likely invalid from a psychometric standpoint. For example, consider the rating of a manuscript's originality using a scale from 0 to 100? A reviewer forced to use such a scale might provide a 70, yet there is no transformative rule that assigns the empirical state of a manuscript a numerical relative. Thereby, it is unclear what a given score of 70 indicates. If such scores factually reflected the translation of a manuscript's scientific value into a numerical representation, then one might use these ratings to form a weighted linear score of such criteria with a predetermined publication threshold score – and fully omit the narrative review. However, given the complete absence of such translation rules, we recommend abstaining from such ratings. Sometimes, online forms require reviewers to complete these ratings; if necessary, one may comment on the arbitrary nature of the ratings in the narrative review.

Broader Perspective on Reviewers' Mission

The remainder of this chapter is devoted to three other classes of recommendations for how to provide a good review. Reviewers should be aware of their impact on and their accountability for the advancement of science; they contribute to shaping format and style of the scientific literature and, by adhering to a distinct set of good reviewer practices, they act as role models (or "influencers") for the academic rules of conduct in the scientific community.

Stimulating and Advancing Science

Beyond their function as monitors and controllers of validity, the peer-reviewing system constitutes a major instrument for stimulating and advancing the growth of science. Peer reviewers determine the rank ordering of the best contributions to a scientific discipline, the allocation of articles to the hierarchy of leading journals and, hence, the textbook representation and the public image of a discipline. That is, reviewers are responsible for the very selection of paradigms and findings that inspire young students and grant proponents, and they co-determine the small subset of topics pursued in future research. *Their impact on the discipline and the identity of a scientific community can be hardly overestimated*.

State-of-the-Art Methodology. In what ways can good reviews stimulate and foster the advancement of science? First, and most importantly, they articulate and administrate the methodology that forms the professional core and the rules of a discipline in a didactically effective way. A distinct methodology is perhaps the most influential defining feature of a discipline that distinguishes professional science from pre-scientific intuition. Therefore, at a more concrete level than logic of science and validity issues, a good review must convey a profound understanding of the underlying methodology in a field. Reviewers' methodological abilities must include the didactic competence to explain and transform methods to every changing field of application.

Novelty and Originality. Second, novelty is a key criterion of fruitful and prospering science. A good reviewer must have a distinct feeling for originality and for the novelty potential of a concept, paradigm, or empirical finding. Moreover, his or her style must convey the clear-cut message that novelty is desirable and productive. Closer inspection shows that novelty in science is a dialectic concept. To understand and to realize vividly what is novel and original, one must understand what is old and long established in the first place. In cumulative science, outstanding contributions are anchored in theoretical priors rather than simply in unexpected results (Fiedler, 2017). Embedding theories creates the potential for compelling innovation and experimental surprises. In any case, novelty is a major theme for a good review and a major dimension of constructive advice.

However, as discussed above, novelty may also derive from the rigorous replication of another novel finding or the test of a novel derivation from a well-established theory, allowing the scientific community to establish the truth or falsity of published research (see above). Reviewers should, thus, not conceive novelty as a narrow concept or confuse it with the provision of new labels; instead, they should, rather broadly, conceive it as insights that would not exist without a given manuscript (e.g., that another published finding is robust and replicable).

Juxtaposing Theories. In a related vein, to understand theoretical innovation, it is often necessary to compare different theories that compete for the explanation of a set of findings. What makes empirical research compelling and outstanding is a diagnostic design that allows for theoretical discrimination in that a focal theory makes a pattern of findings much more likely than competing theories (i.e., yielding a likelihood ratio much higher than 1). Such a critical theory test – *experimentum crucis* – constitutes a stellar moment that scientists are striving for and that, if it is experienced occasionally, is often inspired through constructive reviewing. The ultimate advice that helped researchers to gain a deeper understanding in a truly diagnostic design is often a matter of reframing – viewing a known finding from a new perspective. In any case, good peer reviewing should facilitate such theoretical fertilization.

Transparency. Finally, a prominent goal that became the focus of a new movement, and that a good review will support vigorously, is *transparency*. Good science is inherently public, social, cooperative, open-minded, and striving for cross-validation, explicit documentation, and sharing of data and research tools. Transparency is a good habit of honest and open science, often sponsored by public money. It is also a fertilizer and accelerator of research programs supposed to involve more than one or two labs and a motive for networking, validation, distribution of ideas, and meta-analyses. If the saying is true that "friendship is the cement of science," there is no doubt that transparency and open science are central goals to be propagated in the peer-review process. Transparency is a safeguard against dishonest science and data fabrication, although we deliberately refrain from assigning a good reviewer the role of a police agent or prosecutor.

Cultivating Scientific Literature

The social and behavioral sciences are both empirical disciplines and genres of literature. Cogent findings must be validated, implemented experimentally, and approved statistically and logically. They also must be implemented in the literature through effective communication. This aspect of the scientific game may be called social validation. For example, some of the most influential and groundbreaking examples of psychological science are, to a considerable extent, examples of effectively communicated science. It, therefore, seems justified to conclude that scientific literature and writing style is (almost) as important for the growth of science as validity and good theorizing. Above, we note that many manuscripts may benefit from a higher level of abstraction and formalization. However, a valid and methodologically sophisticated research finding will hardly enter textbooks and curricula if it fails as a piece of literature.

A good example of outstanding research, which was apparently not communicable and never attained the rank that is deserved, is the Zürich model of social psychology (Bischof, 1975) – founded in solid ethological work on incest barriers and fascinating in its theoretical scope. Another example is Egon Brunswik's (1952) underestimated work on probabilistic functionalism. Therefore, a prominent function of the peer-reviewing process, which should not be underestimated, is to foster and improve the communicative, persuasive, ergonomic, and mnemotechnical properties of scientific papers. Likewise, one should not underestimate the extent to which effective peer-reviewing is editorial consulting and training in good writing.

Educating Authors in Good Writing Style. What does this mean, practically? How can this be accomplished? What sort of linguistic or editorial skills can be trained in the peer-review process? Or, conversely, which writing attributes should reviewers focus on to make a research paper easy to understand and memorable? At a very general level, reviewers may look out for good rules of communication, such as Grice's (1975) four maxims of communication. The maxim of *quality* should render all text parts well motivated and credible; according to the maxim of *quantity*, text should be as long and detailed as necessary but not longer, to avoid boredom and fatigue. According to the maxim of *relatedness*, the anaphoric relations between terms and preceding text elements should be unequivocal, such that no terms and phrases remain undefined and unexplained. The maxim of *manner*, finally, is a safeguard against awkward and bizarre language. Another way to explain the Gricean norms is to say effective communication should be cooperative.

Manuscript Organization and Headlining. Next, reviewers may address the organization of a manuscript. An optimally organized article starts with an informative and clearly structured abstract that should mirror the headlining structure of the entire article. The sections or subsections between the headlines should ideally start with an advanced organizer that helps the reader to anticipate the goal and the scope of the next section. Ideally, the same organized contents of an article should reappear in all parts – from the abstract to the theoretical introduction and literature review, the methods and results section, and the final discussion. Good reviews should help to shape articles this way.

Self-Containing Figures and Tables. Another aspect that can have a profound influence on the persuasive power of the entire article is the preparation of technical details provided in figures and tables. Reviewers may check if figures and tables are self-containing; that is, readers should be able to understand all information from the figures and figure captions or from the tables and table headings alone. Self-containing figures are key to strong publications, and a reviewer's task is to teach and coach authors to provide figures and tables, which are also important to break up a paper into clearly visible main parts.

Maxims of Sufficient Contents. Finally, reviewers need to check the sufficiency of the content. The old maxim of providing as many details as required by someone who wants to replicate an experiment is still in force. Substantial information – such as the wording and operationalization of the experimental manipulation, the chief dependent measures, the crucial stimuli, and the main parts of the instruction – should not be hidden in the supplementary materials; they have to be presented in the manuscript proper. The language should be modest and factual rather than sonorous and sexy. The difference between scientific literature and a feuilleton article should be obvious. Good reviewers help authors shape their articles in accordance with these maxims.

Good Reviewer Practices

A good reviewer is not only an inspiring advisor, constructive consultant, trendsetter, and arbiter. Good reviewers are role models who exemplify and illustrate the rules of conduct that enable trust and cooperation among scientists. Although it is not easy, and it may appear somewhat patronizing to formulate obligatory rules of conduct, we believe that the core set of good reviewer practices in Table 33.2 is hardly contestable. We believe that the entire discourse becomes much more motivating and constructive when participating reviewers subscribe to these good-practice rules; they are, of course, not complete. Table 33.2 is rather meant as a checklist or prompt to contemplate and generate a more exhaustive set of good reviewer practices.

We believe the points listed in Table 33.2 speak for themselves. They need not be discussed and illustrated in too much detail. Suffice it to mention that authors' frustration and negative affect associated with peer reviewers, and the resulting lack of motivation and loss of achievement, are often due to the failure to observe these procedural rules. In the worst case, bad reviewing might discourage authors and deter them from daring to conduct original research. Many journal portals and editorials are replete with good researcher practices and author obligations but hardly ever mention what might be called authors' rights and reviewers' obligations.

Table 33.2 *A core set of good reviewer practices, conceived as carrier of academic culture*

Good reviewer practices	Explanatory comment
Avoid personalism	The target or reference of all criticism is the research or manuscript contents, but never the author(s). Reviewers must avoid personally insulting, depreciating, derogatory, or intimidating comments.
Basic author rights	Every author has the right to pursue his or her own research question. Reviewers must not impose divergent research goals and their own standpoints on the authors' work.
Methods pluralism	There are different approved methods to answer a research question. Reviewers must not oblige authors to adopt their own methods preferences.
Provide references and testimony for your critique	Critique must be based on the logic or literature. Reviewers should provide concrete references and evidence when alluding to allegedly existing counter-evidence or overlooked findings.
Be constructive. Suggest solutions and remedies, beyond limitations	Most manuscripts have flaws. In addition to pointing out mistakes, reviewers should suggest solutions, if possible.

Again, the table is not meant to be comprehensive; it is intentionally incomplete, leaving room for the reader's own convictions to supplement our recommendations regarding good practices in the peer-reviewing process. After all, a well-functioning system must have the ability to learn, and the learning function of the peer-reviewing system must be essentially collective. Thus, Table 33.2 can provide a list of prompts to trigger a collective process of learning and contemplation.

However, we want to address two aspects of good reviewing practices that underlie the entries of Table 33.2. The first aspect concerns reviewer style; this is a softer aspect, leaving more room for alternative opinions. The second aspect, namely conflicts of interest, leaves less room for divergent opinions. Let us briefly discuss both aspects before we turn in a final remark to the substantial role played by the editor in the peer-reviewing process.

Style

A good review is not authoritarian and paternalistic – it encourages and convinces authors to follow the provided advice. This is best achieved by avoiding a hostile and confrontative writing style, that only provokes defensive reactions and face-saving techniques. An ideal check on this style is a variant of the golden rule to treat others the way oneself would like to be treated. The abstract notion of a generally supportive, cordial, or polite review tone can be broken down to a few simple, easily understandable guidelines.

Criticize the Work, Not the Authors. It is a different message when "a manuscript contains a mistake" compared to "the authors made a mistake" or when "the manuscript misses a key reference" compared to "the authors missed a key reference." The respective former statements imply a fixable flaw; the latter imply unfixable deficits residing within the authors. Keep in mind that authors identify with their research and with their writings. This is clearly related to our first entry in Table 33.2.

Stay Concrete. The manuscript that missed a reference presents an opportunity for authors. The manuscript that is "sloppy" presents a threat to the self-esteem of authors and a face-saving motive in a defensive revision process. The move from the concrete examples to an overall evaluation (i.e., "the manuscript misses and misspelled many key references" vs. "the manuscript is sloppy") should, if at all, occur at the end of a review. Criticizing authors on an abstract level should be generally avoided (e.g., "the authors are sloppy"). Good reviewers will particularly avoid depreciating remarks about linguistic incompetence. Failure to submit perfect English is either a case for copy-editing or, in more severe cases, reviewers may recommend authors to draw on the help of native speaker (in the case of non-native speakers) or even editing services.

Provide Solutions. If possible, reviews provide ways forward. Instead of stating the fact that a given manuscript misses key references, a good review also provides these references, or good advice regarding where to find them, and may explain why these are key. The same goes for statistical procedures or errors in reasoning. Instead of stating that an analytic choice is problematic, a good review provides the better alternative. Admittedly, there is a threshold of time and investment for good reviewing; in an ideal world, reviewers should share their solutions for the problems they found in a manuscript. They should not dismiss a manuscript simply due to fixable errors.

Beware of Adjectives and Adverbs. Adjectives and adverbs energize a text, but in peer reviewing, they should be used sparsely. Their main function should be differentiation (e.g., “a major concern and a minor concern”) but not emphasis (e.g., “a grave error”). One straightforward method is to simply drop adverbial qualifiers and adjectival attributions and rigorously check if they are absolutely necessary. This approach changes the former sentence to: One method is to drop qualifiers and check if they are necessary. Even when they serve the function of positive evaluations, adjectives and adverbs have little value; editors care less about the assessment that a manuscript is brilliant and elegantly written and more for the reasons for this assessment.

Conflicts of Interest

Editors call upon reviewers, who are experts in their field and who have often themselves made strong contributions to a given field. There are several implications that follow from the reviewers’ investments in a given field.

Reviewers must resist the temptation to pursue their own instrumental motives, for instance, by requesting that their own work be cited or imposing their own subjective views on a manuscript. This is certainly a judgment call. If a reviewer criticizes the manuscript authors’ literature review as insufficient or biased, he or she should provide appropriate references (see Table 33.2) that may sometimes represent reviewers’ own work. However, reviewers should avoid requesting unnecessary citations that serve no other purpose but to increase the visibility of their own work.

Likewise, reviewers might be tempted to favor manuscripts that support their own theoretical or personal view. The most prominent examples are failed replications of existing research, and the authors of the original research are invited as reviewers. The impulse to defend one’s own work and look for errors in the replication attempt is obvious – to be impartial is the basis for good reviewing. A strategy to avoid unwanted consequences of such investments is to distance oneself from one’s own work. Speaking colloquially, if a finding, a hypothesis, or a theory has passed the publication gate, it is alone out in the wild and should thrive or fail without further help from the original authors. Potentially, if the personal investment is too high, a reviewer should even consider declining a review invitation.

Finally, reviewers might be tempted to favor manuscripts by people they are connected with (e.g., by ongoing collaborations, by previous shared authorships, student–teacher relations, etc.). Most journals, and in particular funding agencies, have clear guidelines when it comes to these conflicts of interest. If in doubt, a good reviewer should contact the handling editor or the respective science officer and fully disclose any potential conflicts of interest.

We refrain from any recommendations concerning multiple reviews of the same manuscript for different journals. Although no rational argument strictly prohibits drawing on the same reviewer comments – provided they are valid and cogent – reviewers or editors may believe in the advantage of novel sources. We believe the decision to re-review should be made by individual reviewers.

A Note on Open Access and Alternatives

As we argued above, the critical test of a manuscript is not within the reviewers' hands; it is in the hands of the community. Thus, with the almost unlimited space of the Internet, it is possible to construe an open-access publication model in which authors present their research to the community without peer review, and editorial decisions are placed on openly accessible servers. Readers may then fulfill their critical function of evaluating, replicating, elaborating, or employing the presented research directly.

In an ideal world, such an open-access publication model would be possible and, in our signal detection parlance, we would set the most lenient decision threshold possible. However, one must acknowledge that the scientific output, even of small subdisciplines, has increased to a level that calls for a division of labor. Authors must decide for themselves to which journal they want to submit (e.g., to the most widely distributed or to one more specialized), and the threshold is adjusted accordingly. The reviewers, given the task to judge whether a manuscript is interesting or not for a given audience, evaluate a given manuscript accordingly. Finally, readerships come with certain expectations to the pages of a given journal and scrutinize the presented research accordingly. This division of labor facilitates effective communication of research.

Of course, one may implement a similar system in open-access journals, yet we believe that the necessary structures of uploading, readers as reviewers, and commenting, would copy a substantial part of the existing peer-review system. The main and obvious advantage is that the labor invested by reviewers and editors does not foster the financial gains of large publishing companies – it genuinely fosters the advancement of science. Independent of the underlying system, we believe that the guidelines stated here help to advance the scientific knowledge.

The Editor

In the end, the value and the constructive advice function of even the best review is contingent on the third and most powerful player in the reviewing game – the journal editor (see Chapter 34 in this volume). Editors make the final decision on a given manuscript submitted for publication and function as an arbiter or referee with a substantial, or even determining, influence on the style and quality of the review process (as well as the overarching long-term function of the journal). Still, notwithstanding the power and responsibility of the editor, the quality of good peer reviewing may be remarkably autonomous and independent of the editorial style. Handling editors may exhibit a democratic or authoritarian style; they may contribute their own arguments and opinions to the review process or confine themselves to averaging or combining reviewer votes, and may apply strict or liberal decision styles. Yet, the outcome of the entire process depends, to a remarkable degree, on the selection and the performance of good reviewers doing an honorable job in a goods game with important consequences for science and scientists.

Conclusion

Throughout this chapter, we have adhered to a pluralistic perspective on good reviewing, refraining from restrictive and paternalistic temptations to impose specific ideologies and preferences on a good reviewer (e.g., regarding significance testing, model fitting, online research). We are convinced that such a pluralistic, open-minded attitude is itself essential for an ideal reviewer profile. In doing so, we hope that we have provided some insights that transcend specific research domains, journal types, and scientific trends, and have contributed to helping reviewers to fulfill their mission to the best of their abilities. After all, the role of a reviewer is almost as important for the advancement of science as the role of the original contributors whose work is reviewed.

Acknowledgments

The work underlying the present article was supported by grants from the Deutsche Forschungsgemeinschaft, awarded to both authors (FI 294/30-1, UN 273/5-2).

References

Alves, H., Koch, A., & Unkelbach, C. (2017). Why good is more alike than bad: Processing implications. *Trends in Cognitive Sciences*, *21*, 69–79.

Amabile, T. M. (1983). Brilliant but cruel: Perceptions of negative evaluators. *Journal of Experimental Social Psychology*, *19*, 146–156.

Bischof, N. (1975). A systems' approach towards the functional connections of attachment and fear. *Child Development*, *46*, 801–817.

Brunswik E (1952) The conceptual framework of psychology. *Psychological Bulletin*, *49*(6), 654–656

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*, 193–217.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*(4), 297–312.

Duhem, P. (1954). *The Aim and Structure of Physical Theory* (Translated by P. P. Wiener). Princeton University Press.

Earp, B. D. & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*, 621.

Edlund, J. E., Cuccolo, K., Irgens, M. S., Wagge, J. R., & Zlokovich, M. S. (2022). Saving science through replication studies. *Perspectives on Psychological Science*, *17*(1), 216–225.

Fiedler, K. (2017). What constitutes strong psychological science? The (neglected) role of diagnosticity and a priori theorizing. *Perspectives on Psychological Science*, *12*(1), 46–61.

Fiedler, K., McCaughey, L., & Prager, J. (2021). Quo vadis, methodology? The key role of manipulation checks for validity control and quality of science. *Perspectives on Psychological Science*, *16*(4), 816–826.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (eds.), *Syntax and Semantics, 3: Speech Acts* (pp. 41–58). Academic Press.

Hempel, C. G. & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, *15*(2), 135–175.

Higgins, E. T. (1997). Beyond pleasure and pain. *American Psychologist*, *52*, 1280–1300.

Marsh, H. W. & Ball, S. (1989). The peer review process used to evaluate manuscripts submitted to academic journals: Interjudgmental reliability. *The Journal of Experimental Education*, *57*(2), 151–169.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist*, *38*(4), 379–387.

Quine, W. V. O. (1980). Two dogmas of empiricism. In W. V. O. Quine (ed.), *From a Logical Point of View*, 2nd ed. (pp. 20–46). Harvard University Press.

Rosenthal, R. & Rosnow, R. (1984). Applying Hamlet's question to the ethical conduct of research: A conceptual addendum. *American Psychologist*, *39*, 561–563.

Sturgeon, T. (1957). On hand . . . offhand: Books. *Venture Science Fiction*, 1, 49.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, *1*(1), 1–26.

Unkelbach, C., Koch, A., & Alves, H. (2019). The evaluative information ecology: On the frequency and diversity of “good” and “bad”. *European Review of Social Psychology*, *30*, 216–270.

Unkelbach, C., Alves, H., & Koch, A. (2020). Negativity bias, positivity bias, and valence asymmetries: Explaining the differential processing of positive and negative information. *Advances in Experimental Social Psychology*, *62*, 115–187.

Wagenmakers, E. J., Marsman, M., Jamil, T., et al. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*, 35–57.

Yaniv, I., Choshen-Hillel, S., & Milyavsky, M. (2009). Spurious consensus and opinion revision: Why might people be more confident in their less accurate judgments? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(2), 558–563.

34 Handling Submitted Manuscripts: As Editor and Author

Lisa L. Harlow

Abstract

Submitted manuscripts usually have an arduous journey, while also having the potential to make significant contributions that reach wide and relevant audiences. In this chapter, I offer a path and guidelines for journal submissions; this includes both the editor's perspective on handling submitted manuscripts and implications for the authors. Although journals may vary in how manuscripts are handled, the following three main phases most likely occur in some form: (1) submissions are screened to determine their appropriateness for a journal; (2) manuscripts that remain after screening are usually assigned to reviewers by the editor or associate editor; and (3) manuscripts that remain after the review process are accepted and published. I'm hopeful that the information will be helpful to editors and authors by elucidating the process of handling submitted manuscripts and improving the chances of successful and productive contributions.

Keywords: Manuscript Submission, Screening, Desk Reject, Revise and Resubmit, Editor Tasks and Decisions, Associate Editor, Author

Introduction

If you have conducted research and submitted a manuscript for possible publication, you have accomplished a lot and may wonder what happens next. Three phases are described about the tasks and decisions made to process submitted manuscripts, focusing mainly on the perspective of an editor with some input on the viewpoint of an author. The information that is given for each phase is not exhaustive or new but summarizes several considerations based on the literature and my experience as an associate editor for two journals for six years each (i.e., *Structural Equation Modeling*, and *Psychological Methods*), a journal editor in chief (of *Psychological Methods*) for six years, a (Taylor & Francis/Erlbaum) multivariate applications book series editor for 25 years, and a research author for almost four decades. It is important to realize that the process of editing a journal can involve a number of steps and roles, depending on the journal and the resources available. If there is a large readership and an organization or society that offers support for the journal, associate editors may be assigned to help the editor. Similarly, access to reviewers may vary depending on the size and nature of the journal, with some organizations (e.g., American Psychological Association [APA]) providing a database of reviewers, which could include their areas of expertise, the number of reviews conducted, the average rating of the reviews, and the number of days taken to provide a review. When available, this detailed information on reviewers is very helpful and can facilitate the editorial process considerably.

Realizing that there is variation across journals, there are generally several phases that an editor will undertake: (1) screening initial submissions; (2) assigning manuscripts for review; and (3) accepting manuscripts for publication. Each of these three phases involves tasks and decisions that are discussed in more detail, below, to elucidate the steps that an editor makes and to briefly suggest how the author fits within that phase.

Phase I: Screening Submitted Manuscripts

Journals often have 50 to 1,000 or more manuscripts submitted each year, highlighting the magnitude of the role of the editorial staff. For example, a set of 29 journals published by the APA (APA, 2020a) received from 66 (*Journal of Experimental Psychology: Animal Learning and Cognition*) to 1,045 submissions (*Journal of Applied Psychology*) in 2019. Needless to say, not all submissions are published. Several steps are taken in this phase to adequately screen incoming manuscripts, before a final decision is made.

Step 1: Screening for Plagiarism

Before a manuscript is considered, most journals now check for possible plagiarism using a program such as iThenticate (www.ithenticate.com), Grammarly's Plagiarism Checker (www.grammarly.com/plagiarism-checker), or others. These programs assess the amount of overlap or redundancy that a submitted manuscript has when compared to billions of web pages. The proportion of overlap that signals a problem may vary from journal to journal, although usually an editor would not be happy considering a submission that has 10% or more redundancy with an existing paper, not counting the references. As an editor of *Psychological Methods*, I would send back a paper that had 10% or more overlap, particularly if the redundancy was with the introduction or discussion from one or two specific articles, including the author's previous work. If there seemed to be just isolated phrasing that was similar to a number of articles, especially if it was in the methods or results sections, I would give the author(s) a chance to restate with novel content and resubmit. If the next submission had little or no overlap with other articles, I would proceed with Step 2 screening.

Step 2: Assessing the Relevance of a Manuscript

The second step involves close attention to the mission statement of a journal. A manuscript can be very well written, but if it is not consistent with the mission of a journal, it will not be received well and will need to be redirected.

Editor's Initial Considerations for Submitted Manuscripts

At this second step in the initial screening phase, an editor will want to verify whether a submitted manuscript addresses the mission of the journal. A mission statement is usually placed on a journal's web page and provides input on the main focus of the journal and the kinds of articles that are most likely to be published in the journal. This section of the web page may also have sample published articles that are featured, as well as editorial statements by a current or previous editors on specific topics of interest and the type of studies that are welcomed. The editor will read a submitted manuscript with a particular eye on the title, abstract, hypothesis tests, research questions, analyses, results and discussion, and references. An editor knows whether a paper is in line with the journal's mission and, if it is not, will desk reject it (i.e., reject the manuscript without sending it out for review).

Authors' Initial Considerations When Submitting a Manuscript

Authors have a better chance of making it through initial screening steps if they carefully read a journal's mission statement and objectively ask what type of paper they are preparing. Even if it is a very good paper, it may not be received well if the journal does not emphasize that focus (e.g., theoretical, empirical, applied, quantitative, qualitative, etc.). Not enough researchers appear to take this step seriously – it is extremely important. One question for an author to consider is: Who is the intended audience (specific researchers, the broader field)? Some journals speak to a particular readership, whereas others want to reach a very wide range of researchers. Look at the leading journals in the discipline, based on metrics such as the Altmetric Attention Score (i.e., how much attention articles get from, say, Twitter or blogs), the number of times articles are downloaded, the CiteScore (i.e., the number of times an article is cited in the last three years), and the impact factor – the average number of citations for journal articles published in that journal during a designated time frame, usually two to five years.

Check journal metrics, but don't necessarily just go with the highest values. A paper has to be consistent with what the journal publishes. To clarify the focus, type key words into Google Scholar to find journals publishing on this topic. A set of articles that addressed that topic will be listed, showing where they are published. If there are a couple of papers listed from the same journal, it may be a good place to consider. If there are none published in a journal of interest, it may not be a good match. However, if there are a great deal of papers on a specific topic, it may be that the article's contribution would not be novel enough for that journal. It is also worthwhile to see the references cited in relevant articles and where they are published. Authors should check several journal options and decide. In sum, if the author(s) did a good job convincing the editor that the manuscript is consistent with the journal's mission, it can be considered for the third step in the screening process.

Step 3: Assessing the Contribution of a Manuscript

Manuscripts that make it to the third screening step are evaluated based on whether they make a meaningful contribution to the larger literature.

Editor's Considerations for Verifying a Meaningful Contribution

It is not enough to just discuss a topic mentioned in the journal's mission; editors will want to convince themselves that a paper will make an important impact on the field. Albertine (2010), writing in the journal *Experimental Biology*, agrees. As an editor, I saw a number of excellent articles that appeared to discuss a relevant topic. However, I also had to weigh whether I thought the article met the standards of the journal and offered more than an incremental contribution.

Authors' Considerations about the Manuscript Contribution

Authors can help editors to see a manuscript's contribution by clearly describing it in an author submission letter. The letter should state how the paper matches the mission of the journal and adds to the literature over and above what already published articles provide. Before submitting, authors should make sure that there is a good headline to a paper, in the title and abstract, and that they have selected a pertinent journal; this could involve making sure that other articles have addressed their topic in that journal. Then, they should find a place to cite these papers and add the relevant references to the manuscript. Whereas it is not mandatory to cite articles from the journal to which one is submitting a manuscript, it is good scholarly practice to be inclusive about relevant publications. A thorough check of other articles also helps solidify the conclusion that an article offers something more than other papers on this topic. It is also likely that seeing one or more cited articles from the submitted journal will help in convincing an editor that a manuscript is in alignment with others published there.

Step 4: Assessing the Quality of a Manuscript

If a manuscript covers a relevant topic that could have some promise for making a contribution, the next step is ensuring that the research has been conducted and conveyed adequately.

Editor's Considerations for Assessing the Manuscript Quality

An editor will want to see how well a study was designed and conducted and evaluate the quality of the writing. If a study is poorly designed, it has little chance of being considered for publication. However, some editors may be less critical if the study appears sound, but the writing could still use some work. Some journals could recommend that an author check with an editing service, particularly if English is required and is not the first language of the author. On this point, Lake (2020) argues that a paper with poor writing is easier to correct than one that is not designed well. Pierson (2004) agrees that poor writing quality may not be the main reason for rejecting a paper, but it could still lessen an editor's perception of the merit of the manuscript. When combined with other perceived deficits (e.g., in analysis or interpretation), a paper that is not well written may be more likely to be rejected.

Authors' Considerations about the Manuscript Quality

Authors should take a lot of care in conducting their research and writing up the results. The manuscript needs to be very clear and accessible, speaking to the larger issues in the field, while conveying the main points of their research so others understand what was done and possibly how they could apply it themselves. In this vein, Raitskaya and Tikhonova (2020) point out that authors whose first language is not English tend to face greater desk rejections and more revisions if the manuscript gets past an initial screening. The authors caution that researchers who are submitting to an international journal need to be aware of global perspectives on their research topic; the submission must be relevant within a larger theoretical framework than may be found within a single country. They also suggest having a manuscript reviewed by someone else before submitting to a journal, particularly if English is required at a journal and is not the native language of the author(s).

Making a Screening Decision about a Manuscript

If the editor determines that a manuscript has successfully made it past the initial screening steps, the manuscript will usually be sent out for review, which is discussed in the Phase II section below. However, if the editor decides that a manuscript does not make it through all four steps, the screening process stops, and it is desk rejected – the paper is rejected by the editor without being sent out for review (e.g., Lake, 2020). To this point, Lake (2020) claims that one of the main roles of an editor is to decide on whether to desk reject an article or send it out for review. As the editor of *Research in Nursing & Health*, she found that only a third of the submitted articles were ready to be reviewed – the remaining submissions were desk rejected without review. LaPlaca et al. (2018a), in the field of industrial marketing management, estimate that 85% to 90% of the submissions across most fields end up being rejected. Mendiola Pastrana et al. (2020) speculate that many of the rejections in specialized medical journals are due to a lack of understanding about the review process. There may be some truth to this supposition, as recent literature suggests that many (if not most) journal submissions are desk rejected. The amount varies depending on the focus of the journal and the number of researchers trained in that area.

A recent review of desk rejection in 33 political science journals found that 5% to 76% (40% on average) of the submissions were desk rejected (Garand & Harman, 2021). Similarly, during my term as editor of the journal *Psychological Methods,* an average of 50% of the submissions were desk rejected (Harlow, 2017). This may seem somewhat harsh or dismissive, but it is usually based on thoughtful considerations of whether a paper is a good match and would make a substantial contribution to a given journal.

Fortunately, desk rejections usually occur within about two weeks so that an author has timely notice to consider another course of action for their paper. Judge (2008) even reported that the journal that he edited, *Corporate Governance: An International Review*, would take no more than four days to issue a desk reject. This is not always the case, however, as desk rejections may not happen until several weeks or longer after a submission; Teixeira da Silva et al. (2018) label these as tardy rejections that can cause unnecessary and frustrating delays for authors and journals. Below are considerations from both an editor's and then an author's perspective on what is involved for a desk rejection.

Editor's Considerations with a Desk Reject

One of an editor's main responsibilities is to ensure high standards and success for their journal without over-burdening reviewers who volunteer their limited free time for peer review. For some journals, the editor or associate editor may also be volunteering their time; this heightens the need to be expeditious about the initial screening of manuscripts. To that end, it is often imperative that editors be the first line of screening to determine which manuscripts to assign to reviewers for possible publication and which to send back to the authors without sending out for review. Knowing the mission of the journal and the current state of the field, an editor is in a good position to make practical initial decisions that can save authors and reviewers a great deal of time and help maintain the integrity and focus of their journal. Readers also benefit from effective desk rejections, as they seek out and come to expect a specific kind of article from a particular journal. When a submission does not match the mission of the journal, no matter how well constructed it may be, it won't find the audience it is intended to reach – it needs to be redirected to provide the opportunity for a more relevant stream of dissemination.

Although journals may vary on how they initially evaluate submissions, there does seem to be some consistency in what is expected and, conversely, what is not acceptable. For example, in a survey of over 30 major journal editors in political science, Garand and Harman (2021) found that more than 90% agreed on three categories of reasons for desk rejection – inconsistency with the journal's mission, lack of a clear contribution, or poor quality. That said, Garand and Harman caution that having a large number of desk rejections may not always be viewed favorably by the field, arguing that there are also good reasons to let reviewers decide, so as to avoid having just a single perspective on the outcome of the manuscript. There is not an easy solution, as sending out every submission for review could deplete the reviewer pool and discourage reviewers from participating. Still, the reality of having an abundance of journal submissions, with a somewhat limited pool of experienced reviewers, means that desk rejections will most likely continue to be a strong possibility for editors and authors.

To help ameliorate the rejection of an otherwise reasonable paper, I often recommend to the author(s) one or more journals that might be receptive. These alternative journal options are made based on familiarity with the publishing mission of other similar journals or by checking the Internet to see what other journals tend to publish papers in that particular area. It is of note that I have even been sent an occasional thank you note for helping an author redirect their manuscript – and doing so within a short time of their submission. Below is a brief example of a decision letter for a desk reject, without sending the paper for review.

Dear Dr. xxx,

Thank you for submitting your manuscript to our journal. With the large number of submissions that we get, there is a need to screen submitted manuscripts to gauge how well a manuscript is consistent with our editorial mission (*see below). This process is essential in saving the time of authors and reviewers so that papers that are not a good match can be redirected to a more pertinent outlet.

Given my assessment, I regret that I am not sending your paper out for external review and am not inviting a revision or resubmission. Although I appreciate the potential value of your manuscript, the focus is too narrow and the paper would not make an incremental contribution to our journal, which is very competitive.

It is possible that a more specific journal would be interested in your manuscript. For example, you might consider the xxx journal, or the journal, xxx. The former journal welcomes papers on topics similar to yours, and the latter has published papers using a similar theoretical framework. Either of these or a similar journal would reach a more relevant and potentially more receptive readership than would be the case for our journal, given your focus.

As a minor point, the author guidelines for writing style should be consistently followed when submitting to a journal that requires this format.

In closing, I thank you for your interest in our journal and wish you all of the best in your ongoing work, which could reach an appropriate audience at a more applicable journal.

Sincerely, xxx, Editor

* [Insert a copy of mission statement for the author]

Authors' Considerations with a Desk Reject

It is never easy to receive news that a manuscript was rejected without even being sent out for review. However, authors should know that they are in good company when this happens. This kind of feedback usually occurs fairly soon after a submission so that there is time to focus on whether or how to redirect the paper. If the editor conveys that the paper has serious design flaws or does not appear to make a contribution, an author may have to put aside the possibility of resubmitting the manuscript – it is unlikely that the paper is worth revising and submitting elsewhere. This will no doubt be disappointing, but it is good to remember that most researchers have more research ideas than they have time to complete the projects. Letting go of one that does not seem ready is fine.

However, if there appear to be concerns that an author can correct, but the former editor did not give an option to resubmit to that journal, it is possible to refocus the paper in a more meaningful direction and resubmit the paper to a different journal. For example, if the feedback is that an article simply did not fit within the journal mission, an author could choose to continue working on it and still have a good chance of getting the paper published if they find the right journal. In this regard, Lake (2020) offers good tips that lessen the probability of a manuscript being rejected without review. Suggestions include checking the scope of the article to see if it is line with the journal, plainly articulating the contribution of your study, and going over the writing – particularly of the abstract, methods, and results (Albertine, 2010).

Foremost, it is important not to get discouraged. It takes time and experience to know which journal seems like the best match for a particular research article. Somewhat akin to picking the correct answer in a multiple-choice test, you can't just pick the option that mentions the general topic (e.g., organizational behavior, political science, psychology, sociology). It is important to choose a journal that has the kind of audience that would value a specific manuscript.

Authors should check journal submission requirements and follow them very closely. At a minimum, authors should have a clear and concise submission letter and format the manuscript according to the journal's author guidelines (Johnson & Green, 2009). Mention how a manuscript exemplifies one of the publishing priorities of a journal, and make sure that the body and references of the paper follow the author guidelines. See, also, if there are limits on the number of pages, tables, figures, or word count. Having an incorrect format is a direct signal that the paper was not fully intended for that journal or that you didn't research it enough to verify the requirements.

It is also good to have someone review a paper for readability before (re)submitting. However, an author should not wait until a paper is perfect – it never is. It may come as a surprise that the master painter, Leonardo Da Vinci, prepared a number of drafts and never actually considered that he had finished the Mona Lisa in his lifetime (Smith, 2020). Although it was a portrait commissioned by a local aristocrat, Da Vinci apparently did not consider his work ready for general viewing and continued to work on it for 16 years. Although we may not have a research equivalent of the Mona Lisa, there is a point at which the paper is good enough for submission. A good thing to strive for is to make sure that a friend or relative would be able to understand the paper even if they are not in that field of study. A paper that is coherent and accessible is much more likely to get reviewed positively and have a wider impact.

Having reconsidered a manuscript carefully, have a backup journal or two in case of being rejected from the top choice, and don't get disheartened. J. K. Rowling was rejected by 12 publishers before her Harry Potter books began getting accepted (Hall, n.d.). Luckily, she thought it was still worthwhile to continue to work on her ideas until the 13th publisher was receptive to her first book, showing the benefit of not giving up on your project if it still beckons you. More practically, Pierson (2004) makes a good case for sending out your manuscript and then moving on to your next project. A highly published colleague of mine would also have agreed, often reminding me that it is better to have a manuscript sitting on the desk of an editor or reviewer than stalled on your own desk.

Phase II: Assigning Manuscripts for Review

If a manuscript makes it past the initial screening phase, it is usually assigned to an associate editor who sends it out for review. The associate editor generally selects two or three reviewers that have expertise in the topic of the manuscript, after making sure that the selected reviewers have provided reasonable reviews in the past and have not been over-taxed with too many review requests. A good guide is to limit the total number of requests for a specific reviewer to no more than two or three per year. If an associate editor finds that a reviewer appears to be tapped to review 4–10 manuscripts per year, it would be reasonable to suggest to the editor that the reviewer be added to the editorial board or possibly considered as an associate editor if there is an opening for another one.

Another consideration is whether a journal encourages authors to suggest one or more reviewers for their paper; and, conversely, to suggest individuals who the author would not recommend as reviewers – providing a brief reason for each. In my role as editor, I rarely select more than one, if any, of the author-suggested reviewers to lessen the possibility of reviewer bias. Of course, if it became difficult to find a reviewer for a manuscript, possibly because the topic was rather specific, I might seek out a recommended reviewer; however, it is possible that the review could be somewhat less objective than one from a reviewer that was not suggested by an author. In contrast, I would almost always honor a request to avoid a specific reviewer, particularly if the author stated that the individual had a competing view that might possibly interfere, if only inadvertently, with an unbiased evaluation of the opposing manuscript.

Even with a good journal pool of reviewers and a possible list from the author, it can be difficult to recruit reviewers. The editor or associate editor will need to stay on top of the situation, particularly if a reviewer never responds to the request and it does not seem likely that a review will be forthcoming from that individual. To avoid protracted delays, it is a good idea to have a backup list of alternate reviewers to call on if one from the original set turns you down or simply doesn’t respond. Although most reviewers are fairly responsive and can complete a review in a reasonable time, I sometimes have had to go through a set of 6–10 individuals before finding two reviewers for a manuscript. In that situation, it may be that the topic of the paper was too specialized or did not appear to have enough interest for potential reviewers to accept the offer to review. It may even be that it is so difficult to find a reviewer that a paper that the editor initially thought had passed through the screening steps would still have to be rejected without review, although this is rarely the case. The deciding factor is whether the editor comes to think that the inability to find a reviewer is due to the inappropriateness of the manuscript, or the simple unavailability of the specific reviewers that were selected.

After successfully finding two or three experts who agree to review, the associate editor or editor needs to set a time limit on when the review is due. This can vary across journals, ranging from several days in a rapid-turn-around journal, to a more common three-to-four-week time frame. When all of the reviews are returned, it is time to make a decision.

Making a Decision on a Reviewed Manuscript

A manuscript that has just been reviewed is not automatically accepted for publication. Usually one of two scenarios can occur, and each are described below.

Scenario 1: Rejection of a Manuscript with No Option to Resubmit to that Journal

Most journal submissions are rejected, either by the editor alone or after reviewer input. A report from the APA Council of Editors showed that 36% to 91% (with an average of 71%) of the manuscripts submitted to APA journals in 2019 were rejected (APA, 2020a). Along these lines, LaPlaca and colleagues (2018a) surmise that 80% to 95% of all submissions to leading journals are rejected. These large percentages highlight that getting published takes much effort and does not usually happen easily. This is good to realize so that researchers can go into the process fully aware of what is involved, whether from an editor's or an author's perspective, and then take steps to avoid rejection. Mendiola Pastrana et al. (2020) and Pierson (2004) provide useful input on how to lessen the chances of being rejected from a journal in the health sciences; the suggestions could apply to other kinds of journals as well. For example, they highlight the importance of checking the mission of a journal, spending time considering and describing the design, methods, results, and discussion, and being careful about the writing. Let's consider how a rejected manuscript can be handled from the perspective of an editor and then an author.

Editor's Considerations with a Rejection and No Option to Resubmit

A journal editor oversees all submissions to a journal and is responsible for making the initial assignment of manuscripts for review, as well as the final decision to reject, revise and resubmit, or accept a manuscript for publication. When the confluence of reviewers' input and the editor's reading of a manuscript point to the need to reject a manuscript that does not appear appropriate for a journal, the author is informed that their paper was rejected and is not invited to prepare a revision. This can be distressing news for an author, who must now decide how best to move forward. Fortuitously, such a rejection is usually accompanied by a good deal of input from the editor and reviewers, which provides specific points on how the paper was evaluated. Below is a brief example of an editor's letter that is rejecting a reviewed manuscript with no option to resubmit to that journal.

Dear Dr. xxx,

Thank you for submitting a revision of your manuscript. I can see that you and your coauthors put a great deal of thought and effort into this manuscript, and that it addresses an important area of study. That said, I agree with the two reviewers and associate editor, who provided very thoughtful and extensive comments to convey that there is not enough justification or detail in the paper. The work sounds too preliminary for our journal. Much more analysis would need to be conducted to provide clear guidelines for readers.

Given that the topic has some merit, you might consider another journal (e.g., xxxx, xxxx, or xxxx) and see if one of them has a close match between their mission and the goals of your manuscript. Although I cannot predict how well received your manuscript would be at these journals, in any of these or a similar journal you would reach a more relevant audience than would be the case for our journal with your paper.

I hope that you find the comments shared by the reviewers and associate editor helpful as you decide how to proceed. I do not always see such in-depth feedback on journal reviews and believe that it is worthwhile to consider the suggestions offered as you seek out another journal.

In closing, I regret that I do not have better news and want you to know that only a small percentage of submissions are actually published in our journal. Further, I believe that you will find a relevant home for your paper in another journal.

Sincerely, xxx, Editor

Authors' Considerations with a Rejection and No Option to Resubmit

When an author has a reviewed paper rejected, it stings. Usually, anywhere from one to even six months or more could have passed since submitting the manuscript; this may create a slow building of hope that the paper would be accepted. After getting over the disappointment, an author should go over the reviews very carefully to see the specific points that were raised by the editor and reviewers. There are almost always insightful ideas on how to improve a paper before revising it and resubmitting it to another journal. A word of caution is that, whereas an author doesn't have to attend to every criticism – especially if not resubmitting to the same journal, there is still a need to address as many points as seem reasonable to improve the chances of success at another journal. This is especially important as reviewers often provide feedback for a number of journals, and it is not uncommon for a researcher to be asked to review a manuscript that they just reviewed – and provided considerable input for – at another journal. Ignoring thoughtful feedback, particularly when it could substantially increase the likelihood of getting published, could certainly antagonize reviewers and lead to another missed opportunity.

In addition to reviewer comments, authors can check other sources that could offer useful strategies to consider. For instance, Johnson and Green (2009) give possible solutions to common errors when submitting a paper along with an excellent table that lists resources delineating checklists and guidelines, depending on the research focus. The APA has a publication manual (APA, 2020b) that gives input on every step of the process of preparing a manuscript for publication in psychology and other social and behavioral science fields. LaPlaca et al. (2018a) also provide sound suggestions for publishing effective articles in top journals.

Other articles provide guidelines for specific circumstances. Appelbaum et al. (2018) and Cooper (2018) offer definitive sets of standards to implement when submitting research that uses quantitative methods. Levitt et al. (2018) give standards for publishing articles using qualitative inquiry, mixed methods, or meta-analysis. Stroup et al. (2000) also provide guidelines for publishing articles using meta-analysis of epidemiological studies. Holmbeck and Devine (2009) prepared a checklist for articles concerned with developing and validating measures. Kennedy (2018) gives basic information on publishing more broadly in health-based journals. Still another guide is offered by Degele (2010), who provides information about how to submit ScholarOne manuscripts.

The bottom line is if the feedback offers any hope that the paper could potentially be improved and sent to another journal, authors should do so. This may involve collecting additional data and reanalyzing the data. Otherwise, it is probably best to put aside a rejected paper if it is given stark criticisms that do not appear surmountable. Although this is not a comfortable situation, most of us have experienced this kind of rejection and can move past it to focus on more promising papers from our work. We can also learn by taking steps to avoid the kind of criticisms related to a rejected manuscript that did not have an option to revise and resubmit. My first journal submissions, which came as a result of a great deal of effort and thought, were simply not as focused and sound as I might have initially thought at that stage of my career. Practice and experience definitely help fine-tune one's focus and writing, providing a greater probability that future research submissions will have a better reception.

Scenario 2: Rejection of a Manuscript with the Option to Revise and Resubmit

An editor's decision letter that a manuscript has been rejected but a revision and resubmission is invited is truly good news – for both the editor and author. It means that some precious time was devoted to reviewing a manuscript by the editor and usually several reviewers, and the majority view is that the paper shows promise. It is not a guarantee that a revision of a paper will be published, but there is at least a reasonable chance of publication if the authors thoroughly attend to the comments made by the editor and reviewers.

Editor's Considerations on a Rejection of a Manuscript with the Option to Revise and Resubmit

Contrary to what some authors may think, an editor usually wants to offer an author the opportunity to revise and resubmit their manuscript for possible publication. An exception is when a manuscript is viewed as high risk for rejection due to having a number of apparent deficiencies, even if there appears to be some promise. In this latter case, a manuscript may be rejected if reviewers indicate that the problems outweigh the potential contribution, or the editor may let the author know that their manuscript has a high-risk status of being rejected even if they are given an opportunity to revise and resubmit. For most submissions, however, if an editor is at this decision point, it is usually because the manuscript was viewed as successfully passing the initial screening and reviewer processes and has at least some chance to make it into publication. This is encouraging, as the review process almost always offers ways to improve the quality and impact of a manuscript and is very worthwhile (e.g., DeMaria, 2011). Thus, in this scenario, the editor and reviewers are usually expecting that the author will submit a rigorously revised manuscript that will hopefully make its way to publication.

Below are brief highlights of an editor letter with an offer to revise and resubmit.

Dear Dr. XX,

Thank you for submitting your manuscript for review and consideration for publication in our journal. I appreciate the opportunity to review the manuscript. I have now received three reviews of your manuscript from experts in the field and am able to make an editorial decision at this time.

Reviewers provide excellent input, although there was wide variability about whether the paper should be published here, which made it difficult to come to a conclusive decision. One of the reviewers suggested rejection, one suggested acceptance and one suggested a revision. Whereas I can see the points that each of the reviewers raised, I am leaning toward asking for a major revision as I believe that your paper is consistent with the mission of our journal and has the potential to make a contribution if the details and writing are clarified. If you agree to revise and resubmit, please attend to the concerns presented below by each of the reviewers. As the reviewers provide very detailed and helpful input, I will not reiterate their points here. I believe that a manuscript that attends to the reviewer input, including extensive suggestions by Reviewer 1 and other informative comments from Reviewers 2 and 3, could provide a very accessible and readable manuscript on this issue, which is consistent with our mission and audience.

If you decide to revise the work, and I hope that you do, please submit a list of changes or a rebuttal against each point which is being raised when you submit the revised manuscript. We request that your revision arrive within four weeks. If for some reason you find that you cannot meet this deadline, please contact us as soon as possible in order to make other arrangements. To submit a revision, go to the journal website and log in as an Author. You will see a menu item called Submission Needing Revision and will find your submission record there.

Sincerely, xxx, Editor

Authors' Considerations on a Rejection of a Manuscript with the Option to Revise and Resubmit

When a paper is rejected with an invitation to revise and resubmit, the author has been given a golden opportunity. The reviews will have a wealth of advice on how to make the manuscript more accurate, accessible, and readable (see: Lovejoy et al., 2011; Steer & Ernst, 2021; Su'a et al., 2017; Tikhonova & Raitskaya, 2021). My main advice is to follow up on each and every one of the editor's and reviewers' comments in a thoughtful and carefully conducted revision; then, document the changes in a detailed author response letter that is thorough, polite, and convincing (see Williams, 2004). Below is a brief example of how to address several different kinds of comments from an editor in an author response letter that accompanied a revised manuscript:

Dear [insert editor name],

Thank you for the opportunity to resubmit this manuscript. We appreciate the comments from you and the reviewers. Below, we address each comment and hope that you find the revisions satisfactory.

Editor comments:

1. The manuscript could use more explicit grounding in theory to support the hypotheses and conclusions.

Author response: We agree that the manuscript could use more theoretical grounding and have added two paragraphs on pages xx to xx, which discuss our theoretical framework that serves as the basis for our initial research questions and later interpretations of our results.

2. Reviewers 1 and 2 questioned whether the sample size was adequate for the study that was conducted and would like to see a power analysis to see if there was enough evidence to detect the effects that were expected from the hypotheses.

Author response: Thank you and the reviewers for the suggestion to conduct a power analysis. We did so by using the open-source program, G*Power (Faul et al., 2009), which indicated that we needed at least a sample size of 119 for our study. Thus, we described in the method section (on page xx) that our sample of 120 participants would provide at least 95% power to detect a medium f2 effect size of 0.15, which is equivalent to an R2 effect size of 0.13 in the multiple regression analysis with our three predictors of the single outcome variable (Cohen, 1988).

References:

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Erlbaum.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behavior Research Methods, 41, 1149–1160.

3. Reviewer 3 recommended that you consider conducting a multilevel model with your data, using your two continuous independent variables as level-one predictors, and the dichotomous grouping variable as a level-two variable. This would allow an assessment of whether the two predictors were significantly related to the outcome, after taking into account whether the data were nested within the two different geographical areas.

Author response: Thank you for the suggestion. My co-authors and I discussed this option and concluded that the limited number of only two geographical areas in our current study would most likely not be sufficient to adequately conduct a multilevel model on our sample of 120 participants. Further, we believe that the current study still provides reasonable evidence that geographical area is important as we found a significant medium effect for this dichotomous variable. We would like to plan a future study in which we would collect data from at least 20 geographical areas with a larger sample, which would provide a sufficient number of groups and participants to more reliably assess whether our participants were significantly nested within a larger set of geographical areas and whether the two predictors still showed meaningful effects on top of a geographical area effect. We mention this plan in the paragraph on future directions in our discussion section on page xx.

Sincerely, xxx

As shown in this example, not every editor or reviewer suggestion has to be adopted; although, if not, there should be a compelling reason why it wasn't (e.g., DeMaria, 2011). In the author response example, the first comment was readily addressed, as it was an important request from the editor to more clearly articulate the theoretical framework for the study. Similarly, the second comment was based on input from two of the reviewers and appeared important to address by conducting a statistical power analysis to verify the soundness of the sample size and effects. The third comment was said to be from only one of the reviewers and asked for a statistical procedure to be considered that would not be optimal with the study's limited number of two groups – when data from at least 20 groups should ideally be available for a multilevel model. Thus, the author response politely indicated that this option could be considered for a larger future study, but the authors thought that the current study would not allow an adequate test of such a model; however, they still provided reasonable initial evidence on a limited geographical effect.

Regardless of the extent and nature of the editor and reviewer feedback, it is almost always a good idea to take an editor up on an offer to revise and resubmit. The journal has already invested their time and expertise in helping to shape a submitted paper into a better form. Although an author can rebut a point made by an editor or associate editor, it is important to accept the input graciously and take the time to address or refute each of the comments and document the changes clearly and completely in the revised manuscript (e.g., by highlighting changes) and in the author response letter. Helpful input on revising manuscripts can be gleaned from seasoned colleagues and from input in the literature (e.g., Altman & Baruch, 2008; DeMaria, 2011; LaPlaca et al., 2018b; Pierson, 2016; Price, 2014).

Phase III: Accepting a Manuscript for Publication

The last phase of handling submitted manuscripts involves less work and occurs less often. There is minimal uncertainty at this point, although there are still a few details to address as an editor or author. A reasonable estimate is that about 15% to 25% of author submissions will end up being accepted (Forsyth, 2021;

Harlow, 2017), although there is a range depending on the nature of the journal and number of submissions. With high-impact journals that receive a lot of submissions, the percentage of accepted manuscripts is usually small. Nonetheless, the few that reach this point are usually well worth the effort that went into them; however, the timeline from submission to publication can vary widely from a several weeks in some fields, such as cardiology (e.g., Kusumoto et al., 2021), to 21 months in an initial review of over 4,000 studies in the field of biomedical research (Andersen et al., 2021).

Editor's Considerations for Accepting a Manuscript

When an editor sees that the requested revisions have been satisfactorily completed and the (majority of the) reviewer recommendations are to accept the paper, the author will usually be sent a decision letter that the paper is accepted or conditionally accepted upon completing any remaining minor revisions or publication forms. Editors and journals vary as to how many rounds of revision authors will need to make and the percentage of authors that reach this point. It is common to request multiple revisions before a paper is accepted. A study of revisions and citations for the journal *Business, Management and Accounting* found that the number of revisions that were made of a manuscript was directly related to the number of citations it later received (Rigby et al., 2018). Given that finding, it is to everyone's benefit to ensure that accepted manuscripts have been adequately reviewed and revised. Below is a brief example of a letter from the editor, offering the author acceptance after minor revisions.

Dear Dr. xxx,

Thank you for sending the second revision of your manuscript to our journal. I was the only one to review this version of the manuscript as it has already been reviewed twice by reviewers.

I appreciate the careful attention to the comments from me and the three reviewers on the previous version. The revision was much clearer and had more justification for the conditions that were included. I also liked that the manuscript was reorganized and shortened somewhat by moving some material to appendices, making the manuscript more readable. Thank you again for depositing all of the computer codes, as well as the raw data, into an Open Science Framework repository. This is helpful to readers who wish to follow up on aspects of the manuscript of interest to them.

Here are a couple of other comments:

I did not see a response to Reviewer 1's point 6 (i.e., "I would suggest to add a small paragraph summarizing the main results; alternatively, a table could be used to summarize the main trends"). Please provide input on this point.

A translational abstract is needed that describes the essence of the study in very clear and understandable language, and that is not as technical as the original abstract.

Please go over the manuscript one last time to make sure it is concise and accurate, including the reference section.

Pending receipt of a minor revision that takes the above three comments into account, I am conditionally accepting the manuscript for publication in our journal.

Sincerely, xxx, Editor

Authors' Considerations with a Conditional Acceptance

If an author is fortunate enough to have been offered conditional acceptance if designated revisions are made, they are in a very good position. Most likely, they have done a great deal of work to get to this point and deserve a lot of credit. When an author receives a conditional acceptance, it is very important to make the last revision(s) and complete the requested publication forms in a timely manner, so as not to hold up the publication process (DeMaria, 2011). This kind of acceptance rarely happens after the first submission of a manuscript and usually takes one or more rounds of review – at least some of which are by reviewers – with the editor possibly making the final decision without reviewer input after the last revision(s). After this point, authors can legitimately celebrate and list the manuscript in their curriculum vita as "accepted." Great job!

Conclusion

The main editor decisions include: (1) a desk rejection after screening a manuscript – the editor decides that a paper will be rejected without sending it out for review; (2) agreeing to send a manuscript on to an associate editor or set of reviewers; (3) a rejection based on input from reviewers and associate editor – the editor decides that there is not enough merit to invite a resubmission of a revised paper to that journal; (4) a rejection based on largely promising comments from the reviewers and associate editor – the editor offers an invitation to revise and resubmit a paper to the same journal; or (5) an acceptance or conditional acceptance pending the completion of any final steps (e.g., signing forms, proofing the final manuscript). The editor is active in each of these decisions, although the first decision involving screening initial submissions usually takes most of the editor's time, with the remaining decisions largely driven by feedback from reviewers and an associate editor.

For authors, the chance of having a manuscript eventually published in a specific journal increases if they make it past the first screening phase, and on to the reviewer and possible revision phases, even if it takes several revisions. Still, authors should realize that the largest percentage of submitted manuscripts are rejected without review (i.e., desk rejected). Conversely, the smallest percentage (almost none) of submitted manuscripts are offered acceptance after little or no revision. In between

https://doi.org/10.1017/9781009010054.035 Published online by Cambridge University Press

these extremes, reviewers are involved and, if an author has a strong manuscript, there is some probability that they will be invited to revise and resubmit their paper; this is a very good outcome.

In sum, the art and science of submitting an article and following it through to revisions and publication takes practice, patience, and persistence (LaPlaca et al., 2018b). Editors, associate editors, reviewers, and authors have travelled a long way to get to the point of having a manuscript accepted. No one expects sports players or musicians to be perfect the first time they make an appearance. Further, they are admired and rewarded for showing great determination when they vigorously practice – often daily. For some reason, researchers don’t always realize that the need for persistent practice also applies to them.

Although there is no sure formula for handling submitted manuscripts and getting published, it is important to seek opportunities to develop expertise in best practices. Those who would like to serve as an editor can volunteer to review more and to inquire about serving as an associate editor if it appears that a position could be available. Authors should follow the guidelines in their field and those for a specific journal to maximize a favorable response, and make sure to offer and articulate a unique and valuable contribution in each paper that is submitted. They should take time with writing as the best papers have a wide reach when they are clear and understandable. Rejection is to be expected – the large majority of submissions get rejected.

In the best outcome for editors and authors, an invitation is made to revise and resubmit a paper, increasing the likelihood of publication and impact. Most of all, the process should be enjoyed. Thoughtful and insightful advice from an editor, along with compelling and effective research by authors, requires inspiration and effort and is always worth it in the end – no matter how many revisions it takes to be published. Everyone benefits when a research project is carried from the initial draft through to purposefully and tenaciously following up on the submission and revision process. Fortunately, the processes of screening and reviewing submitted manuscripts, along with possible revising and resubmitting by authors, become easier and more rewarding the more that experience is gained and the longer that editors and authors are involved in research. Happy submissions!

Acknowledgments

Thanks are offered for support from Grant G20RR030883 from the National Institutes of Health (PI: D. H. DeHayes), to the American Psychological Association for their support during my terms as associate and chief editor of one of their journals, and to colleagues, students, and family who have helped encourage me in my journey toward meaningful scholarly work.

References

Albertine, K. (2010). Advice for submitting manuscripts to scientific journals. *Experimental Biology*, *24*(S1). https://doi.org/10.1096/fasebj.24.1_supplement.8.1

Altman, Y. & Baruch, Y. (2008). Strategies for revising and resubmitting papers to refereed journals. *British Journal of Management*, *19*, 89–101. https://doi.org/10.1111/j.1467-8551.2007.00542.x

Andersen, M. Z., Fonnes, S., & Rosenberg, J. (2021). Time from submission to publication varied widely for biomedical journals: A systematic review. *Current Medical Research and Opinion*, *37*(6), 985–993. https://doi.org/10.1080/03007995.2021.1905622

APA (2020a). Summary report of journal operations, 2019. *American Psychologist*, *75*(5), 723–724. http://dx.doi.org/10.1037/amp0000680

APA (2020b). *Publication manual of the American Psychological Association*, 7th ed. American Psychological Association.

Appelbaum, M., Cooper, H., Kline, R. B., et al. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board Task Force report. *American Psychologist*, *73*, 3–25. http://dx.doi.org/10.1037/amp0000191

Cooper, H. (2018). *Reporting Quantitative Research in Psychology: How to Meet APA Style Journal Article Reporting Standards*, 2nd ed. American Psychological Association. https://doi.org/10.1037/0000103-000

Degele, L. (2010). New illustrated guide for authors submitting to ScholarOne Manuscripts™ (formerly Manuscript Central). *Editors' Bulletin*, *6*(2), 40–42. https://doi.org/17521742.2010.516195

DeMaria, A. (2011). Manuscript revision. *Journal of the American College of Cardiology*, *57*(25). 2540–2541.

Forsyth, A. (2021). Peer review in a generalist journal. *Journal of the American Planning Association*, *87*(4), 451–454. https://doi.org/10.1080/01944363.2021.1958551

Garand, J. & Harman, M. (2021). Journal desk-rejection practices in political science: Bringing data to bear on what journals do. *PS: Political Science & Politics*, 54(4) 676–681. https://doi.org/10.1017/S1049096521000573

Hall, D. (n.d.). J K Rowling turned down by 12 publishers before finding success with Harry Potter books. Available at: https://riseupeight.org/jk-rowling-harry-potter-books/.

Harlow, L. L. (2017). The making of *Psychological Methods*. *Psychological Methods*, *22*(1), 1–5. http://dx.doi.org/10.1037/met0000141

Holmbeck, G. N. & Devine, K. A. (2009). Editorial: An author's checklist for measure development and validation manuscripts. *Journal of Pediatric Psychology*, *34*(7), 691–696. https://doi.org/10.1093/jpepsy/jsp046

Johnson, C. & Green, B. (2009). Submitting manuscripts to biomedical journals: Common errors and helpful solutions. *Journal of Manipulative and Physiological Therapeutics*, *32*(1), 1–12. https://doi.org/10.1016/j.jmpt.2008.12.002

Judge, W. (2008). The screening process for new submissions. *Corporate Governance: An International Review*, *16*(6), i–iv.

Kennedy, M. S. (2018). Journal publishing: A review of the basics. *Seminars in Oncology Nursing*, *34*(4), 361–371. https://doi.org/10.1016/j.soncn.2018.09.004

Kusumoto, F. M., Bittl, J. A., Creager, M. A., et al. (2021). High-quality peer review of clinical and translational research. *Journal of the American College of Cardiology*, *78*(15), 1564–1568.

Lake, E. T. (2020). Why and how to avoid a desk rejection. *Research in Nursing & Health*, *43*, 141–142. https://doi.org/10.1002/nur.22016

LaPlaca, P., Lindgreen, A., & Vanhamme, J. (2018a). How to write really good articles for premier academic journals. *Industrial Marketing Management*, *68*, 202–209. https://doi.org/10.1016/j.indmarman.2017.11.014

LaPlaca, P., Lindgreen, A., Vanhamme, J., & Di Benedetto, C. A. (2018b). How to revise, and revise really well, for premier academic journals. *Industrial Marketing Management*, *72*, 174–180. https://doi.org/10.1016/j.indmarman.2018.01.030

Levitt, H. M., Bamberg, M., Creswell, J. W., et al. (2018). Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 26–46. http://dx.doi.org/10.1037/amp0000151

Lovejoy, T. I., Revenson, T. A., & France, C. R. (2011). Reviewing manuscripts for peer-review journals: A primer for novice and seasoned reviewers. *Annals of Behavioral Medicine*, *42*(1), 1–13. https://doi.org/10.1007/s12160-011-9269-x

Mendiola Pastrana, I. R., Hernández, A. V., Pérez Manjarrez, F. E., et al. (2020). Peer-review and rejection causes in submitting original medical manuscripts. *Journal of Continuing Education in the Health Professions*, *40*(3), 182–186. https://doi.org/10.1097/CEH.0000000000000295

Pierson, D. J. (2004). The top 10 reasons why manuscripts are not accepted for publication. *Respiratory Care*, *49* (10), 1246–1252.

Pierson, C. A. (2016). The four R's of revising and resubmitting a manuscript. *Journal of the American Association of Nurse Practitioners*, *28*(8), 408–409. https://doi.org/10.1002/2327-6924.12399

Price, B. (2014). Improving your journal article using feedback from peer review. *Nursing Standard*, *29*(4), 43–50. https://doi.org/10.1080/00131857.2020.184651910.7748/ns.29.4.43.e9101

Raitskaya, L. & Tikhonova, E. (2020). Seven deadly sins: Culture's effect on scholarly editing and publishing. *Journal of Language and Education*, *6* (3), 167–172. https://doi.org/10.17323/jle.2020.11205

Rigby, J., Cox, D., & Julian, K. (2018). Journal peer review: A bar or bridge? An analysis of a paper's revision history and turnaround time, and the effect on citation. *Scientometrics*, *114*, 1087–1105. https://doi.org/10.1007/s11192-017-2630-5

Smith, R. (2020). Learning from Leonardo – The importance of the rough draft. Available at: https://blog.shrm.org/blog/learning-from-leonardo-the-importance-of-the-rough-draft.

Steer, P. J. & Ernst, S. (2021). Peer review: Why, when and how. *International Journal of Cardiology Congenital Heart Disease*, *2*, 100083. https://doi.org/10.1016/j.ijcchd.2021.100083

Stroup, D. F., Berlin, J. A., Morton, S. C., et al. (2000). Meta-analysis of observational studies in epidemiology (MOOSE): A proposal for reporting. *JAMA*, *283*(15), 2008–2012. https://doi.org/10.1001/jama.283.15.2008.

Su'a, B., MacFater, W. S., & Hill, A. G. (2017). How to write a paper: Revising your manuscript. *ANZ Journal of Surgery;* *87*(3), 195–197. https://doi.org/10.1111/ans.13847.

Teixeira da Silva, J. A., Al-Khatib, A., Katavić, V., & Bornemann-Cimenti, H. (2018). Establishing sensible and practical guidelines for desk rejections. *Science and Engineering Ethics*, *24*, 1347–1365. https://doi.org/10.1007/s11948-017-9921-3

Tikhonova, E. & Raitskaya, L. (2021). Improving submissions to scholarly journals via peer review. *Journal of Language and Education*, *7*(2), 5–9. https://doi.org/10.17323/jle.2021.12686

Williams, H. C. (2004). How to reply to referees' comments when submitting manuscripts for publication. *Journal of the American Academy of Dermatology*, *51*(1), 79–83. https://doi.org/10.1016/j.jaad.2004.01.049

35 Grant Writing Basics

Tamera R. Schneider, Howard C. Nusbaum, and Jennifer N. Baumgartner

Abstract

This chapter is for *all* academics, from students and faculty to professional staff at research centers and institutions. The content draws upon our experiences from when we were budding scholars, to experienced scientists, and now administrators, including time spent at federal funding agencies. Our aim is to provide information to scholars so that you can write more competitive grant proposals and secure greater resources for your research and scholarship. First, we provide a broad overview of what to consider before you embark upon writing a proposal. Then, we discuss areas for consideration in writing the proposal itself. Finally, we share steps to consider after you have received feedback about your proposal. We also provide some detail about particular funders, including support for international scholarship. As with all scholarship, persistence, collaboration, and support from colleagues are helpful for successfully securing external funding.

Keywords: Grant Writing, External Funding, Scholarship, Grants, Resources, International

Introduction

All research and scholarship require resources. These resources include people, time, infrastructure (e.g., internet, libraries), materials, facilities, and equipment – to name a few. Although some scholars proactively seek out funding more than others, securing such resources can increase your capacity to engage in scholarship. Additional funding can provide various resources, including the critical time needed to devote to your research, specific materials, participant remuneration to collect data for research projects, or student time and training for the next generation of scholars. Securing additional funding can support your capacity to conduct the research itself; it can be an important part of your promotion and tenure case and is often a requirement. Promotion and tenure typically require proficiency in research and scholarship, depending upon the institution's mission. Securing external funding for your work provides a clear signal to your employer that the ideas and methods offered are valued beyond that of the academic institution. Scholars often learn quickly whether their program of research or area of scholarship will require ongoing funding, such as access to specific populations or equipment. As with any area of research or scholarship, a deep grasp of the discipline and an entrepreneurial spirit will be helpful in the quest for securing additional funding for your scholarship.

Before You Get Started

Determining Whether You Should Seek Funding

How do you know if you should put the time and effort for your scholarship toward grant writing? Throughout your academic training, it should become clear whether you need special materials or personnel that require financial support for your research. When one of the authors was in graduate school, she learned that if a primary research area is a less fundable niche curiosity, it is best to have an additional line of research that is more clearly fundable. You will likely gain a sense of whether you need to secure additional funding for your work early during your academic training.

Does your research involve special equipment, tools, or software? Equipment requires regular maintenance and software often requires annual licenses. If the source of resources is unclear, professors and mentors can be a great source of such information. A curious student who is engaged in understanding the scholarship process will be more entrepreneurial in securing resources. If you learn that securing external support is relatively rare in your chosen discipline, finding the means to obtain funding can put your career on a novel and exciting path. If you are in the humanities and arts, for example, could you assist with art projects (e.g., at a local school or after-school program) in exchange for materials? A brief synopsis of your idea and a small budget and justification could help to cement such an arrangement – more on that later. There are many avenues for securing research resources and funding. Engage with potential providers to find overlapping and mutual benefits.

Once you determine that you should or that you want to write grant proposals, the topic of when to start trying to secure such funding in your career may also be clear – generally, the earlier the better. With research questions in hand, but scant pilot data or depth of expertise, early-stage scholars might set their sights on early wins by applying for smaller amounts of funds that are set aside for pilot projects and/or conference travel. Some such funds are often available through student organizations, professional organizations, or from within an academic institution through internal funding competitions. Although it is generally best to have pilot data for a particular project, there may also be smaller awards from federal and other funders that target opportunities for collecting pilot data. For these smaller amounts of funding, the competition may not be as steep. Practicing grant writing through early-stage competitions can provide early lessons on grant writing to different audiences – and maybe actual funds.

Engaging in Research Projects and Questions That Are Likely to Receive Funding

Having a sense that you need funding to support your scholarship, or knowing that you want to obtain additional funding, how do you engage in a fundable scholarship line? Research questions can come from a variety of sources, such as the domain of research conducted in a faculty advisor’s laboratory, compelling coursework posing interesting unanswered questions, new questions offered by ongoing research, conversations with colleagues at professional meetings, and so on (see Chapter 3 in this volume). Is your particular question fundable? It is helpful to have a colleague or mentor who is familiar with your research area and who will know what the current funding priorities are. One of the authors wrote a pre-doctoral fellowship proposal that, given the nature of the hypothesis, the mentor thought would be declined. The proposal was well written and addressed an unanswered question, but it was not timely. Still, there were many good grant-writing skills learned from that endeavor and many successes before and after that experience. The point is that knowing what your funding audience wants to fund is important.

Although mentors can be helpful in knowing what is likely to be funded in particular areas, there are general areas of inquiry that are more likely to be funded. Several research and scholarship areas that will be important for decades to come include climate change, public health, and artificial intelligence. Determining ways that your scholarship can help to understand and address issues relevant to these domains will put you in a better position for securing external funding. Further, you can set your sights on becoming engaged in cross-disciplinary research related to understanding and addressing these domains (see Chapter 32 in this volume). You may be able to leverage your ongoing research program and reframe it to create overlap with different colleagues and available funding opportunities. For example, one of the authors conducted research on persuasion. She received funding to conduct persuasion research in the domain of health disparities and cancer with cancer practitioners. This persuasion research expanded to include understanding and addressing people’s engagement with the environment and climate issues in partnership with municipalities – both of which garnered internal and external funding support.

Beyond staying abreast of what is happening in the field, another way to learn about current and near-term funding is to attend to what governments set as their priorities – from local municipalities to the broader reaches of government. Paying attention to what elected officials state as their priorities will often shine a light on new funded initiatives. These priorities are often in public documents, including the budget; the budget can also provide language that can guide ideas for research efforts as well as content that should appear in proposal writing. Such details point to what these particular funders are seeking to support.

Finding Internal and External Funding Sources

There may be internal funding competitions within your organization that can typically be determined quickly by contacting the research office. These funds are typically limited, as their goal is often to offer seed funding, to launch and strengthen research or scholarship to then secure external funds. For example, seed funds can provide the resources to conduct pilot research, that garners supporting evidence for larger proposals submitted to external funders. Even here, seed funding might be targeted to particular scholars (e.g., early-stage) or particular areas (e.g., climate change), making it important that you understand what the funding priorities are for these internal competitions.

Resources for many types of research and scholarship can be sought from a variety of external sources – government agencies, philanthropic foundations, industry partners, and government municipalities, to name a few. Oftentimes, these entities will put out requests or calls for proposals in areas that they have an interest in understanding or addressing. Other times, there may be standing requests for proposals that originate from scholars' interests. In addition to funding basic research, where new knowledge is generated but the immediate impact of the project is not known, there is often external funding available to target societal interests that have public impact. Public impact research emerges from and feeds back into basic research. Although funders may focus on either or both – public impact research and/or basic research, some disciplines have access to more types of funding (e.g., science, technology, engineering, and mathematics [STEM]) than others (e.g., humanities and arts), and many countries have differing funding priorities.

One way to identify the availability of different types of external funding is to use grant prospecting software tools – these tools exist at most institutions. The library, research office, or research foundation, if there is one, can point scholars to the tools available. Moreover, they often offer worthwhile training for researchers in the use of the particular tools available at the institution. For example, prospecting tools allow researchers to engage in searches facilitated by keywords. Keywords can be related to a general research idea or program (e.g., "public health" and/or "health disparities" and/or "children"), a particular type of funding (e.g., philanthropy or government), particular types of awardees (e.g., early-stage, mid-career), or other ways to narrow the search. Depending upon your interest, entering keywords will bring up specific opportunities or requests for proposals from a variety of funding agencies. The search can narrow or broaden depending upon keyword use or type of funding sought. Such searches can also help you to reconsider your research question by helping you to understand what funders are prioritizing in that domain.

Other great informational sources are colleagues and professional organizations. Within professions, the type of funding typically received by top scholars becomes clear through acknowledgments at professional conferences and through publications. Professional conferences allow you to engage directly with colleagues, their scholarship, and often external funders. Professional organizations may also provide their members with their own internal funding programs and timely updates about discipline-relevant external funding opportunities through a newsletter.

One might also learn about funding opportunities by perusing country-level funder websites; often, you can sign up for newsletters and announcements from such agencies. In the United States, the National Science Foundation (NSF) funds basic research, ranging from social, behavioral, and economic sciences to physics and engineering; the National Institutes of Health (NIH) funds research with biomedical and public health outcomes. Some funders, such as the NSF and NIH, fund a standing portfolio, and they may also set aside additional monies for other funding priorities – often determined by the appropriating agencies' mandates. Both the NSF and NIH send out announcements regularly, and you can sign up to receive these announcements on the funder website.

Websites are also useful for international funding agencies. For example, the Belmont Forum (www.belmontforum.org) includes over 50 countries across six continents, which focus on funding international transdisciplinary research on environmental change. They have a useful website that discusses their purpose, highlights calls for proposals, and provides helpful information about the impacts of projects that have received funding in the past. Philanthropic foundations (e.g., Templeton, Spencer) also engage in targeted funding. If the funder has a website, it will be very helpful to spend quality time there, as the site will convey both funding opportunities that are available and the mission and priorities of the funder. Your proposal should speak to the priorities of the funder and any particular call for proposals that you are responding to with your proposal. In addition, when there are unforeseen crises (e.g., pandemics or natural or manmade disasters), additional funding opportunities often become available from cities, states, regions, countries, philanthropies, and others, for projects that understand and address the urgency.

Early-Stage Writing

Writing a Compelling Proposal

With a general project idea and a funder in mind, you are ready to begin the writing process. Proposals are more competitive when they are explicit about the contribution and compelling need for the project. They do not assume that the proposal reader will take the time and effort to figure it out; such proposals help the reader to understand all aspects of the proposal. Does the project contribute to the field, create transformative knowledge, or address particular concerns of the funder? If a project contributes to the field, but the funder is interested in some practical applications without regard for a field, then you may have a proposed project-funder mismatch. Be sure that you are proposing something that will be relevant for the funder given its mission, a particular call for proposals, or whatever it is that you are responding to for that funder.

It is important to incorporate funder priorities into the proposed project to make it more competitive. For example, whereas the NSF tends to focus (although not exclusively) on contributions to basic science and transformative scientific thinking, the NIH tends to focus on contributions to understanding health; this can include, but is seldom exclusively directed at basic science and novel or emerging treatments. The US Department of Defense (DOD) has different categories of research, which span from basic science to the development and implementation of technical solutions. Different philanthropic agencies (e.g., Gates, Templeton, Robert Wood Johnson) all have different foundation goals and missions. Understanding funder goals and projects that they funded in the past will increase your awareness of the broader audience for the proposed project. Furthermore, when grant writers are explicit about the contributions of a proposed project, it is easier tell whether there is a match between the purpose of the proposal and that of the requests from the funder – whether general or targeted toward a specific domain. The clarity of the purpose of your proposal may take shape after multiple revisions; be patient and give yourself time to write a good proposal.

Project aims that are narrowly focused within a discipline, or that address a small, old, or over-studied problem, will be viewed as too incremental to be competitive. Project ideas that are current and innovative will be more competitive. For many agencies, it may be important to both generate new knowledge and provide "transformative" results. A transformative project can be one that changes the scientific thinking in a field. This type of project may present new strong evidence against the general theoretical assumptions in the field or produce a new theory that accounts for prior research while making new and unique predictions. A transformative project may change other fields – either because they become unified due to the results of the research or because new theoretic analyses yield insights that change the way results are understood. Such a transformative project could lead to new treatments for disease, new engineering solutions, or new social policies. Whether a project proposes to test and reject accepted theories, generate new theories or perspectives on a field, integrate knowledge across fields, or lead to new therapies, educational approaches, or engineering solutions, program administrators are interested in grant projects that will have the potential for an important impact on science, medicine, education, engineering, or society.

Research projects that involve scholars across disciplines to address a problem area from different disciplinary perspectives can be transformative. There are often government or philanthropic funds that are dedicated to address major societal issues or perplexing research questions. When a team of scholars from different disciplines (e.g., social and behavioral sciences with other STEM disciplines, and art and humanities, education, or policy) come together to discuss, understand, and address these issues, the proposed projects are often exciting and groundbreaking (see Chapter 32 in this volume). We encourage you to meet colleagues from different disciplines and different institutions to focus on these major challenges so that we can come to helpful and innovative solutions for the nation and world. These types of collaborations tend to generate novel hypotheses and have the capacity to transform fields and start new emerging fields. Such proposals generate greater enthusiasm from reviewers and those who manage funding portfolios.

Writing to Your Audience: Reviewers and Funders

Having some idea about the project you want to propose and the type of funding opportunity you will seek, you can begin the proposal drafting process. Be sure to keep your audience in mind as you begin the writing process, and consider the writing level before you submit your proposal. Reviewers may or may not be from your discipline. You can imagine that these are educated people, but they may not be conversant in the nomenclature of your field. Write clearly so that the reviewers can understand what the problem is, how you are addressing it, the method or methods by which you will address it, and how your findings will meet the goals of the funder. Generally, the proposal should be written in language geared toward a thoughtful undergraduate, someone who is intelligent but not in the field. There may be other detail needed in particular places where you can show your expertise; there, too, consider that the reader may not be in your field. Avoid too much jargon that leaves your reader blind to what you plan to do and why. Your goal is to write a clear and compelling proposal – a story that gets your readers excited about the project and its potential impacts.

Writing to the Solicitation or Call for Proposals

Some funders, such as the NSF, have published criteria against which proposals are reviewed, and reviewers are expected to evaluate proposals accordingly (NSF PAPPG, 2022). (The NIH also has guidance for its reviewers (NIH 2022).) In some cases, there will be particular goals or priorities that are reflected more precisely in calls for proposals, beyond general program descriptions. These may be called requests for proposals (RFPs), broad agency announcements, solicitations, collaborative research actions (CRAs), and so on, depending on the funder. The program description or RFP will often point explicitly or implicitly to the criteria by which proposals will be evaluated. Be sure to read these carefully and ensure that you have all the criteria addressed in your proposal. If you are unclear about the criteria or the fit of your project with the request, you should contact the program director, often identified in the RFP. If you are unclear about program fit, it would be helpful to write a one-page summary of the project, send it to the relevant program director, and request a phone conversation to discuss its relevance to the program. Program directors typically have a good sense of where the field is going and what reviewers will find compelling, so having a discussion with them, months before the submission deadline, about your well-constructed one-pager can be very helpful in creating a more competitive proposal for submission. See more below on contents of a one-page summary.

One example of a request for proposals is the 2020 call for CRAs by the Belmont Forum (2020). Not only has the administrative office offered the written CRA, but they have also made available a video for proposal writers. Furthermore, they provide information about projects that have been funded in response to prior-year CRAs. Engage yourself with these materials before you start writing to ensure that you can target your project to what the funder seeks to support and that you have a good idea of the criteria required for securing funding.

Setting Up a Support Network for Pre-reviews

Plan to arrange for preliminary feedback about the proposal well before it is due. You might share with colleagues who have been successful at getting grants funded – mentors, your department head, or others – that you are developing a proposal for submission to a particular funder. There are differences of thought about sharing proposals, so if you decide to ask a funded colleague for a copy of a successful proposal and they decline, do not take it personally. Proposals reflect intellectual property. Seek pre-reviewers that you trust and who will agree to keep the project in confidence.

Ask a couple of colleagues if they are able to read your proposal in advance of submission. Ideal colleagues are those in a similar discipline who received funding from the program. Consider those who are in an adjacent discipline and received funding from the program or agency. Ask colleagues whether they could provide a timely pre-review so that you can strengthen your proposal. Expect that co-investigators will provide timely feedback and co-write the proposal to some degree. You might also enlist graduate students or even family members, who are not in your field, for readability. Keep in mind that reading a proposal well and providing thoughtful feedback takes many hours, so engage your support wisely. Generally, you would give a person at least two weeks to engage with a good, near-final draft of the proposal, to fit within their already engaged workload. Generally, you should give yourself at least a month to consider their comments in the proposal before submitting it. Many research offices provide support to grant writers by investing in and finding others to provide confidential reviews. These investments should be utilized well in advance of the due date so that revisions can be made and the outcome more successful. These and similar advance reviews can greatly enhance the competitiveness of a proposal, particularly if there is time to attend to the constructive feedback.

Writing the Proposal

Creating a Coherent Narrative

Writing takes *time* and good writing takes much revision. A poorly written proposal is one that includes too much jargon and/or a haphazard organization; it neither facilitates understanding of the need for the project or its potential impacts. Help your reader by writing the proposal in such a way that you set up the importance of why the project contributes to the funder priorities. The purpose of the project should be clearly weaved throughout the proposal – in the introduction, hypothesis, method, and potential impacts of the project. There should be a common thread such that the introduction sets up a problem tested by the methodology of the project, with impacts that inform ways to understand and/or address the problem. A project is often comprised of multiple studies with different methods. A proposal can quickly fall in rankings when the project is set up so that the first study must succeed to proceed to additional studies proposed. The failing of the first study, for example, should not hinder the need for subsequent studies; all studies should inform different aspects of the problem. The analysis plan, if appropriate, tests the hypothesis so that findings will help to understand the contributions of the project to the field or public impact area, as laid out in the introduction.

Organization of the Proposal

If you are writing in response to a request for proposals, be sure to incorporate the main requirements in your outline before you proceed with writing the proposal. Describe the what, why, and how of what you want to accomplish. When beginning to write a proposal, regardless of its length, there are general areas that should be included, which speak to the what, why, and how. Depending upon additional funder criteria, you might consider the following organization: (1) brief overview, (2) introduction – including the nature of the problem and why this project, (3) the hypothesis, model tested, and proposed study/method, (4) the analysis plan, (5) the team needed to conduct the work, (6) the timeline, and (6) the budget and its justification.

The brief overview is a one-page summary of your project. This one-pager is essentially the specific aims page for US NIH proposals and the project summary for NSF proposals (see below). It can serve as the prelude to the proposal and may be required. It can also guide a conversation with program administrators to discuss program fit. There are four broad areas to cover in the overview, and specific funders may have additional required elements. You should be clear about (1) the problem that the proposal addresses (importance, urgency), (2) why this proposal is best positioned to address it (theory, approach, convergent team, expertise), (3) how the project addresses the problem (hypotheses, methodological approach), and (4) the implications of the project outcomes (societal impacts).

Space is tight, and the overview should pull the reader in by beginning with a compelling statement about the problem that the proposal addresses. This includes its importance and urgency; these should be geared toward funder priorities. When discussing why this proposal is a great fit to address this problem, briefly point to the theoretical underpinnings of the project and the novel approach employed to address the problem. If there are multiple disciplines involved, or if special expertise is required to conduct the project, these should be mentioned briefly. Given a problem, and that the proposed approach is set up to address it well, a statement of the overall hypothesis of the project should be conveyed. If there are multiple studies, there may not be space to mention them all in the one-pager, but the overall hypothesis should be stated clearly.

After providing a brief overview, toward the middle of the one-pager, depending upon the project and number of studies, in a paragraph or so the proposal may outline the method or methods that will be used to investigate the problem proposed by the project. With multiple studies, if there are similar broad-level hypotheses that can be constructed, or similar methods, it may be helpful to denote that for the reader explicitly. For example, if the first three experiments have a similar hypothesis, and the latter three have a similar hypothesis, you could group those into two sub-hypotheses. Furthermore, if Experiments 1 and 4 have similar methodology, 2 and 5 are similar, and 3 and 6 are similar, you could briefly provide clarity about hypotheses and methodologies that way.

There are many different ways to organize the brief overview and proposal – the point, in this paragraph, is to be clear and help the reader to understand what your proposal involves, including denoting hypotheses and your methodological approach. Lastly, clarifying the implications of how the proposed project addresses the problem will leave the reader with a clear idea about the relevance of the work to the intended target. If the funds are meant to serve the public good or come from taxpayer dollars, then conveying the societal impact of the project is important.

The one-pager is often the first piece you work on, and takes a lot of time to develop. It guides your work on the larger proposal; it must also be refined over time. It may also be the last piece you work on, to ensure it aligns with what has evolved as you developed the proposal.

Moving beyond the one-pager, the rest of the larger proposal is a more detailed embellishment of the brief overview. Similar to a manuscript, the writing of a proposal has an hourglass shape; the narrative starts out broad in scope, narrows where one builds the argument and addresses how the investigation will unfold, and then ends broadly in clarifying contributions and impacts. The sections of a proposal generally include: (1) a broad introduction, (2) a discussion of the nature of the problem and why the project is well poised to address the problem, (3) the model tested and methodology of the proposed project, (4) a presentation of the team, (5) plans for analyzing or inferring outcome success, (6) a timeline of the project, and (7) the budget and justification that is clearly linked to the proposed work. The length of the sections varies, and the length also depends upon the total length required by the funder; the order of the sections may vary, too. Ultimately, you want to end up with a coherent narrative that flows and is easy for reviewers to follow.

The introduction provides a broad overview of the proposal and is usually relatively brief – ranging from one to several paragraphs. The introduction provides the opportunity to expand upon the problem. The expanded problem statement pulls the audience in by creating a compelling need to address the theoretical or real-world problem and a sense of urgency about addressing it through the proposed project. This brief section is where the reader gets the overall sense of the significant gaps in knowledge. It should end with a clear statement of the objective of the proposed project. With a clear objective in hand, the proposal should then present the current understanding of the problem in the discipline, or disciplines, why the problem exists, and what is known about potential remedies. The proposal should delve more deeply into the nature of the problem. This part of the proposal ends with a broad statement about the gaps that should be addressed. Next, the proposal should articulate how this project will fill these gaps in novel and thoughtful ways. The proposal should clearly lay out the innovative approach taken by the project and make the case for why this approach addresses the problem well. This section will spur reader interest in the overarching model that drives the project and detail the methodology.

The proposal becomes most detailed as the model, the general hypothesis (or research question) that address the problem, and the proposed research and methods for conducting the research that test the model are presented. This section points to any pilot studies or particular expertise needed and that have been secured; this demonstrates that the proposed project can be carried out successfully (see the section “Providing Evidence That the Project Can Be Conducted Successfully,” below). Providing information about the team, the qualifications that they possess, which inform their role and the work they will engage in for the project, provides more evidence to reviewers that the project can conducted well. The level of detail continues with a discussion of the planned analyses for testing hypotheses and/or evaluating or assessing outcomes so that you can infer whether the proposed project was effective.

Depending upon the organization of your narrative, the proposal might speak to societal impacts here, or after the timeline, but you should be clear about what the broader impacts of the proposed project will be. Although you have not conducted the proposed research, and will not have an answer, the proposal should consist of introductory arguments and a method for addressing the problem such that the project results in some benefit. Speculating logically about what that benefit is, as it aligns with the funder mission and societal impact, will be important. For examples, the NSF and NIH agencies require a statement about how the work affects the public; this is important given that public taxes are the source of their awards. Below are examples of impacts – these will be unique to each project and are an important part of review criteria. For the NSF, the broader impacts area might discuss how the project applies to areas other than the target discipline, groups other than the target group, trains diverse scholars, solves a pressing societal problem, and/or brings together previously unlinked institutions and individuals. For the NIH, the significance section (often found early in the Background and Significance section) should describe the public health impact of the topic of study, a brief description of what is known, critical knowledge gaps, and how the project will address these gaps.

Most proposals will also include a timeline. The timeline for each of the key areas that were discussed in the proposal, and perhaps additional items if they fit, will provide reviewers with a sense of whether the proposed work could be conducted within the proposed project time period. A brief statement about a timeline might be included in a one-pager, but often not. For a larger proposal, the timeline, if included, might include quarterly (or whatever makes sense for the project) time periods across which the major phases are addressed. These phases range from project development to dissemination. An assessment timeline might be included, especially if there are multiple assessments, and a feedback phase wherein the project requires evaluation and reconsideration of methods. Proposals will include a budget and a budget justification. Some institutions provide assistance with budget development, and it is wise to ask – especially if personnel funding is part of the proposal request. Provide a budget and justification that speak to the work required to conduct the proposed project.

Providing Evidence That the Project Can Be Conducted Successfully

Head off any concerns about whether the proposed work can be carried out. Demonstrate that the project is feasible by presenting findings from prior scholarship. This might include published or unpublished pilot studies to show that successful engagement with, and delivery of key aspects of the work needed are part of the proposer's skill set. There is no need to demonstrate that this exact project can be conducted; in fact, the proposal is offered to support that work, but the reviewers will need assurance that any special skills or manipulations are within the skill set of the investigators, consultants, or key personnel.

One way to show that a project can be carried out successfully is by showing that you have experience with the *type* of project that is being proposed. Demonstrate that you can engage in experimental manipulations that result in the intended effects. For example, if the project requires some type of psychological manipulation, provide pilot data that demonstrate you have effectively done this or a similar psychological manipulation before. If particular statistical methods are proposed, demonstrate that there is experience among the investigators, or hire a consultant for the project. Perhaps a key component of the project involves using a piece of equipment or facilities, recruiting special populations, or employing particular technical expertise. Reviewers will have even more confidence by seeing that the expertise has been applied and relevant outcomes have been shared in conference presentations or publications. If you are not an expert in a certain area, you should propose to hire a consultant who can help you to carry out the proposed work successfully.

Another way to demonstrate that you can conduct a project from start to finish is by showing your audience that you engage with the breadth of the scholarship process – from ideation, to application of scholarship methods, to dissemination. Sharing your research findings in presentations at professional conferences and/or publishing your scholarship in peer-reviewed journals is important for reviewers and funders. Not only will your funder get credit when you present your research – make sure you provide such credit, but your projects will be strengthened by conversations you have with colleagues at conferences or through the peer-review publishing process (see Chapter 31 in this volume). Presenting findings at a conference provides a quick win, but be careful about presenting at too many conferences during a year – it can distract from publishing.

Special Considerations for Different Funders

Below we provide detailed guidance for one funding agency, the USA's NIH. There is much overlap with the general guidance about what to include in a proposal, but each funder will have its own guidelines and requirements. Again, proposers should become familiar with these requirements. We also offer guidance for letters of intent (LOIs) to write a proposal; these are required before submitting some proposals.

A Funder Case Study: The USA's NIH framework

Landing an NIH grant can be a crucial key to launch your career, and can serve as a pathway to important discoveries that can dramatically improve public health and reduce disparities. In general, domestic or foreign, public or private, an non-profit or for-profit organizations are eligible to receive NIH funding. Writing a successful proposal requires careful planning and preparation. The single best advice we can give is to *start early*; this goes for any proposal.

The NIH offers different types of grants. Research training and fellowship grants (T&F series) provide opportunities (including international) to trainees at the undergraduate, graduate, and postdoctoral levels. Applicants are required to develop both a training and research plan. NIH research grants (e.g., R01, R21, U01, K99/R00) support discrete research projects, pilot projects, large institutional collaborations, and business innovations. The NIH also supports program project/center grants (P series) that include large, multi-project initiatives, as well as resource grants to enhance research infrastructure. Although each NIH institute, center, and office (ICO) has their own specific funding priorities, and each grant mechanism has its own requirements (be aware of these ahead of time), we have provided *general* writing recommendations that can be applied to most NIH funding mechanisms. These recommendations are similar to the general organizational information provided above but provide unique guidance relevant to the NIH in particular, as a case study.

The Specific Aims is arguably the most important section because it is the first section reviewers read. You only have one page to gain the confidence of reviewers and convince them that your project should be funded. The key is to be thorough yet succinct. Generally, there is an *introductory paragraph* where you introduce the research topic and capture reviewers' attention. Describe what is known, unknown, and how your project fills a need. In the *second paragraph*, introduce your proposed solution. Describe the what, why, and how of what you want to accomplish. Convince your reviewers that your solution is reasonable and that you and your team are the best people to do the work. If available, describe previous research you have conducted on the topic and how the proposal is a logical and necessary progression of this and others' work.

Explicitly delineate your *long-term goal* (overarching research goal), *hypothesis and proposed objectives* (central hypothesis and the overall goals of the application), and *rationale* (what informed your central hypothesis, and what will be gained from completing this work). Next, describe each of your *aims.* Tell reviewers exactly what will be learned from the project. It is important that your aims are related but independent. The failure of one aim should not jeopardize another. Provide a title for each aim and describe, in a few sentences, your methodological approach and its relevance to your central hypothesis. If you have space, consider offering an alternative hypothesis should the original not be supported. Finally, in your *last paragraph*, take a broad stroke and describe what is *innovative* about your project, your *expected outcomes*, and how it is *impactful* (i.e., how your research will help those in need). With a penultimate draft (the one-pager), it is a great idea to check in with the relevant program officer (PO) to ensure that your aims will be competitive and revise accordingly.

Next, the Research Strategy delineates the technical aspects of the project. There are three key components – significance, innovation, and approach. For most NIH grants, you will also need to address *rigor and reproducibility* by describing how your research design and methods will achieve robust, unbiased results. The research strategy should be organized well, aesthetically pleasing (e.g., visually emphasize key phrases, include helpful figures, etc.), and write clearly. The *significance* section should describe the state of the field as it relates to the aims, the long-term research plans, and investigators' previously published work and/or preliminary/pilot data related to the current topic. In addition, discuss the significance of the expected research contribution. Write from the perspective that your audience is highly intelligent, but they are not in your field.

The *innovation* section should show how the proposed project explores new scientific avenues, bridges existing silos, and will generate new needed knowledge. Lastly, the *approach* should describe the experiments that will address each aim. Generally, provide less detail is in this section relative to a scientific manuscript, but provide enough detail so that reviewers know exactly how you will test hypotheses and that you are using validated methods. Where appropriate, explicitly state why you and your team will succeed in performing these methods, leaving little to no doubt in the minds of reviewers. Visual graphics, such as flow charts to delineate steps, are highly recommended. The approach is typically where proposals lose points, so paint a clear picture of what you plan to do and how you plan to do it. As with manuscript writing, try to anticipate what a less than enthusiastic or skeptical reviewer might say, and seek feedback from trusted colleagues.

What Makes for a Good Letter of Intent?

When funding agencies and foundations announce a new grant program or post a call for proposals, there is sometimes a two-step process involving a pre-proposal. Some funders require pre-proposals that are evaluated before a full proposal is submitted (Stephens, 2013) or sometimes a LOI. In this case, the first step is that the agency/foundation requests a LOI from those who intend to submit a proposal. It is important to read the terms of a solicitation or request for proposals carefully, especially to determine if a LOI is the first step. When there is a LOI or similar requirement, it is critically important to read both how the LOI is used by the funding agency – as there are two general ways they are used – and what the agency expects to be included in the LOI. Pay careful attention to requirements about the LOI length, deadline for submission, and the timeline for the entire grant submission process.

Agency solicitations use LOIs in two different ways, and the difference is usually clear in the solicitation. The first use of an LOI is for agency planning – to anticipate the distribution and nature of proposals that they can expect. The agency is primarily interested in the expertise and disciplines of the principal investigators (PIs), the nature of the research being addressed (one specific discipline, transdisciplinary, or interdisciplinary), and the kinds of methods being used. The agency wants to understand who is submitting proposals and the general topics of proposals – to plan and construct the review process. This may include ad hoc outside reviewers or the assembly of a review panel or both. Regardless, the agency recruits reviewers with relevant expertise to review proposals once submitted.

When it is clear in the call for proposals that this is the purpose of the LOI, there is not much point in attending to the elegance of the writing, the fine details of methodology, or the justification of the proposed research. Furthermore, these LOIs are not typically treated as a commitment on the part of the investigators to actually submit a grant. This may be made explicit in the call. In fact, there are often more LOIs submitted than actual proposals received. Investigators need not fret that they will be thought of poorly if they submit an LOI without following up with a proposal. This first kind of LOI is simple to write, need not be revised extensively, and can be thought of as a placeholder for the agency, without concern that its evaluation will reflect on the proposal itself. The LOI contents are evaluated less relative to the second function of LOIs.

The second form of LOI is much more critical to the grant evaluation process and to the ultimate submission of a grant proposal for the solicitation. This kind of LOI is an intrinsic and important part of the solicitation and evaluation process. This too should be clear in the program description and call for proposals. This type of LOI is used for an initial triage process in which the program administrators assess the LOI for the proposal on several dimensions and use this assessment to decide whether to invite a full proposal submission. To be clear, the LOI needs to provide the information that the investigators want the program administrators to understand about the full proposal to convince the administrators to *invite* a full proposal. In this case, the LOI is a kind of sales pitch to convince administrators that what the investigators propose has substantial merit – to get invited to submit a full proposal. The nature of the LOI sales pitch and the details needed depend upon the specifics of the program call for proposals. These specifics need to be read carefully and understood.

With this second type of LOI, there are general principles to guide the preparation of a good and convincing LOI. First and foremost, the LOI should clearly convince program directors that the proposed research addresses the goals of the solicitation. It is important to spell out the case for the reader that the proposed research grant will advance the program goals. Explain how the research questions in the proposal are related to the program goals – the what and why. Second, the LOI should clearly identify how the proposed research will advance those goals. Explain how the proposed methods will increase understanding of those questions – the how. For most LOIs, this does not mean giving specific methodological detail as would be the case in a full proposal but, instead, identifying the nature of the methodological approach and being clear about why this approach is likely to generate new knowledge. If it is a transformative project, the LOI should briefly explain how the possible outcome of the research may lead to this transformation. Requests for LOIs will likely differ in what is expected (e.g., length, detail, and elements requested). However, in all cases, the LOI should make clear, in a few sentences, why the proposed research is relevant to the goals of the solicitation, why it is likely to generate new knowledge beyond prior research, and how it will achieve this.

It is important to address any criteria that are required as noted in the call for proposals. For example, some LOIs require the identification of a "focus area" – a specific domain or project track that the full proposal would be addressing. A LOI may require a specification of the type of project (e.g., empirical or theoretical, testing humans, non-humans, or specific populations). These details are generally identified as a range of choices the investigators select from. The LOI should make clear that there is sufficient relevant research expertise in the team so that the proposed research can be carried out successfully. Sometimes, this can be done directly by listing personnel, titles and affiliations, or providing curricula vitae. If references are permitted, this can be accomplished more subtly by citing the prior peer-reviewed publications of the research personnel and investigators. Depending on the prescribed LOI length and requested detail, briefly establishing that the appropriate research facilities are available for the PIs can be helpful.

Often, the solicitation will elucidate a structure for LOIs, but sometimes this is left to the investigators to determine, with only a length constraint – as short as a page or as long as several pages. Almost all LOIs require a brief introductory statement on the current state of the field or problem that is relevant to the goals of the solicitation, the overall nature of the proposed approach, and an outline of the methods and expertise. As noted above, it is important to indicate how the research may have a substantial impact. This is similar to the Intellectual Merit and Broader Impact criteria used by the NSF, or the theory of change used by the John Templeton Foundation. Sometimes, each of these can be a few sentences and sometimes a paragraph, depending on the guidelines of the LOI provided in the solicitation.

It is worth stating again that the proposal be jargon free. It is important to remember that the program administrators reading the LOI may not be experts in the field of the investigators, may not know the history of research in the area being proposed, and may not be familiar with the methods intended. Also, they may not know the investigators proposing the research. It is, therefore, critical that the LOI includes clear language with little technical jargon even while providing substantiation that the investigators themselves are experts in the research area.

After Submission

The Review Process

Reviewers and program administrators get excited about compelling and transformative research; they all want to usher in new knowledge in the field, but it is not a quick process. The review process for internal and external proposals can take many weeks to many months. It can be difficult to wait for reviews, but be patient. Funders may provide a sense of how long a typical review process might take or an idea of when they hope to announce awards, but the process takes time. Submitted proposals are first vetted for compliance with the call requirements. Securing external reviews is often the most time-consuming. Reviewers may provide their proposal evaluations by mail or they may serve for several years on a panel. In that case, you typically get feedback from both reviewers and the panel discussion. Depending upon the funder, you may not receive any feedback. You could volunteer to be a reviewer, especially to programs that you intend to submit to, by sending your curriculum vitae to program administrators and noting your interest. This will give you a sense of what makes some proposals more successful. Some review processes may call for volunteer reviewers, but others may rely on internal reviews. Ultimately, the process results in the available funds going to some projects and not others.

There are different types of review rankings or scores. For example, the NSF reviewers each use a five-point scale to provide proposal ratings (Excellent, Very Good, Good, Fair, or Poor). Along with individual reviews, proposals often receive a panel ranking that represents the group consensus of where the proposal fits among four categories: (1) Highly Competitive, (2) Competitive, (3) Not Competitive – Fundable, and (4) Not Competitive – Not Fundable. If you find yourself in the "fundable" range (1, 2 or 3), but did not get funding, it is important to reach out to the program administrator, after a few weeks, to discuss the project and its evaluations. A proposal does not require all Excellent reviews to get funded, but rarely will Fair or Poor reviews be part of the evaluation of a funded proposal.

In contrast, the NIH uses a nine-point rating scale in its scoring system (1 = *exceptional*; 9 = *poor*; whole numbers, no decimals). The normalized average of all study section impact/priority scores constitutes the final impact/priority score. Impact scores range from 10 to 90, where 10 is considered the best. Generally, impact/priority scores of 10 to 30 are most likely to be funded; scores between 31 and 45 might be funded; scores greater than 46 are rarely funded. If you find yourself in the "might be funded" range, it is important to reach out to the program administrator to discern the likelihood of funding. This will be based on the ICO's available funding, funding priorities, and balance within the current grant portfolio.

Contacting the Program Officer after You Get Your Reviews

If you get feedback that the project was funded, congratulations! The project planning, but not the spending, can commence. You should read the reviews and other project summaries to help you to strengthen the project before you start planning. If the project was not funded, the program administrator may be able to help you to understand its strengths and weaknesses beyond the reviews and summaries, if provided. However, you should wait a couple of weeks after digesting the reviews before you contact the program administrator for a discussion. The focus of the discussion is to ensure that your reading of the reviews is aligned with what the program administrator believes is pertinent for the reviews, summaries, and the program.

Do you agree on what the biggest gaps were in the proposal and what made the reviewers most excited? The administrator can share anything of importance that was discussed during the panel that may have been underemphasized in the feedback or provide clarity about a mismatch of emphasis in feedback documents. You should discuss what to prioritize in addressing the comments and can share your ideas about how you might do that. The program director will likely not endorse anything, in particular, but can help you to understand what comments you really should address and which were least important. Asking about that explicitly is a good idea. If you were in the fundable range, you should revise in consideration of that discussion. If you were not in the fundable range, you should work with your institution to get more support for your grant writing efforts. Whatever you do, if the project is not funded, please do not take it personally. There are many great projects but only limited funds – persistence is key.

How to Incorporate Reviewer Recommendations into a Revision

If it is not clear from funder guidelines, during your discussion you can inquire about preferences for addressing feedback in the resubmission. Some agencies prefer a prelude at the beginning of the revised proposal that briefly outlines the changes, but that often takes up limited space. Other funders prefer that feedback be addressed in the proposal, where it is relevant, and to address it directly, without referring to it as a response to reviewer comments in particular. For example, if there were questions about a seeming lack of expertise for a piece of equipment, when mentioning that piece of equipment in the proposal, adding information about user expertise or an added consultant with that expertise would be best.

Conclusion

We hope the above provides a useful introduction to grant writing. There are volumes on the topic. Generally, having a good match of your project to the funder, writing a clear and compelling story, and persisting in these efforts will help to secure funding for your research. As with publishing, such efforts are often met with declines, but we hope that you keep trying. As with all research and scholarship, persistence, collaboration, and support from trusted colleagues will go a long way to your successfully securing external funding.

References

Belmont Forum (2020). Transdisciplinary research for pathways to sustainability. Available at: www.belmontforum.org/cras#pathways2020.

NIH (2022). Review criteria at a glance. Available at: https://grants.nih.gov/grants/peer/guidelines_general/Review_Criteria_at_a_glance.pdf.

NSF PAPPG (2022). NSF 22–1 proposal and award policies and procedures guide, October 4, 2021. Available at: www.nsf.gov/publications/pub_summ.jsp?ods_key=papp

Stephens, D. W. (2013). *Writing an Effective NSF Pre-proposal*. ISBN 13: 978–1493547067. Middletown, DE.

36 Teaching Research Methods and Statistics

Jordan R. Wagge

Abstract

This chapter can serve either as a starting point or a recharging point for instructors who are preparing to teach research methods and/or statistics at the college or graduate level. Using empirical work and experience, I'll discuss the importance of these classes and then provide a list of what I perceive to be the most important recommendations and biggest challenges related to teaching these courses. Because so many classes include projects, I have dedicated half of the recommendations and challenges to the ones that involve student projects. It is my hope that a person reading this takes away a set of ideas and feels empowered to teach these courses well. For all the challenges, teaching these classes can be rewarding for the instructor and transformative for the student.

Keywords: Research Methods, Statistics, Teaching

Introduction

Research methods and statistics are required courses for most students majoring in the social and behavioral sciences. According to Friedrich et al. (2000), 89% of psychology programs at institutions in the United States require coursework in research methodology and 93% require coursework in statistics. Despite their ubiquity, few courses are as controversial or polarized as research methods or statistics. Compared to other courses, for example, students may feel unenthusiastic prior to the start of the semester (Rajecki et al., 2005).

Despite having sufficient content knowledge to teach these courses, faculty members can sometimes feel unprepared and avoid this "service" course; instead, they favor classes that match their research specialization(s). However, research-focused courses provide the foundation for the skills that inform our knowledge in these subdisciplines; when we read a claim in a textbook or prepare a lecture on current trends in our discipline, these things have been made possible by accumulating knowledge through research and scholarship. Still, teaching and learning about the process of research can be unappealing to both faculty and students.

It is easy to empathize with both faculty and students here; even those of us who have lighter teaching loads and heavier research loads engage in research mechanics primarily in the service of answering research questions in our subdisciplines or the supervision of student projects. Very few of us wake up in the morning thinking, "*Today*, I am going to *assign* participants to *conditions* in an *experiment*!" Instead, we wake up (sometimes at all hours) thinking about our research questions, our populations, the change we want to see in the world, the statements we want to evaluate, or the inequities we wish to dismantle. Maybe we can do this *through* research, but the statistics and methods themselves are not the end goal for most people. I would argue, however, that teaching these courses can be a vehicle for engaging faculty *and* students in all these things. It can also be a tremendous amount of fun and a transformational experience for students' development, regardless of their eventual career or academic trajectory.

In this chapter, I pull from experience and the literature related to scholarship of teaching and learning to provide a series of *recommendations* and *challenges* for teaching research methods and statistics. In each of these two main sections, there are six bullet points to consider; the first three points are general considerations; the last three are relevant for courses that include (or might include) a project.

Recommendations for Teaching Research Methods and Statistics

Recommendation 1: Take the Leap – You Should Teach It

Teaching Research Methods and Statistics as a Student

The benefits of teaching research methods and statistics are notable, even if you teach these courses as a graduate student. As a PhD student, you learn more about these subjects while doing your dissertation; this can help you develop ideas. You can also talk about your thesis or dissertation in class, using it as an example and gaining experience with "thinking on your feet" in front of an audience – before dissertation proposals and defenses.

Another benefit is learning about tools outside of your immediate domain. Research labs tend to employ a narrow range of methodologies; in graduate school, I mainly analyzed categorical data gathered from within-subjects psychophysics experiments with a tiny number of participants. I was well trained in some specific experimental techniques, such as block randomization. Still, I was less familiar with a host of other techniques employed by psychological scientists, including between-subjects designs, qualitative research, and survey research. Teaching a research methods course as a graduate student elevated my understanding of these approaches and techniques above my undergraduate-level education; to teach them, I had to understand them.

Another related advantage of teaching these courses as a graduate student is mentoring undergraduate students in research. There are several avenues by which you could do this. First, you could have students in your course present or publish their work. Many research methods and statistics courses feature a project (discussed later in this chapter), for which students collect, analyze, and present data to the class. It might be possible to have students seek institutional review board (IRB) approval prior to data collection and then present their work at a university, undergraduate, or professional conference as a poster or oral presentation. Many institutions have at least one day of the year when they celebrate student research with an institution-wide conference; at my university, this occurs in April and is called “Student Scholar Day.” This is a good place for students to showcase what they’ve been working on and to make the work you’ve done mentoring them more visible to the campus community. The same follows for more regional or national events. Don’t forget to cite the presentation on your curriculum vitae under the heading “Mentored Research”!

A second benefit of mentoring students is that you could have students in these courses help with the research ideas you have for your research program. For example, the student projects in your course could have a common theme related to your work, and mentoring these projects would help you develop your knowledge base. Just don’t forget to give the students credit or invite them to co-author if they think of a good research idea that you use later!

Third, you could develop professional relationships with students who you might then mentor after the semester has ended. Although I will argue that the first two situations are better, any of these can result in presentations or publications for yourself or your student and help with job searches or awards for teaching and mentoring. I also found that mentoring students in research as a graduate student helped establish a lab later; this could also be very helpful for future work in industry related to managing teams. If you are enthusiastic and encouraging in your teaching, this will help you recruit students to work with you or commit to presenting/publishing their work after the semester has ended. Plus, being mentored in research is an excellent experience for students.

Finally, and perhaps most importantly to some, teaching research methods and statistics will likely make you more marketable on the job search – in industry or academia. As previously mentioned, it is a course that not every faculty member wants to or does teach, yet I argue that, for several reasons, it is a course that is well served by having a full-time faculty member teaching it. First, bringing in relevant examples from your research can help garner students’ enthusiasm and investment in the course. Second, these are courses that often require assistance outside of class, so having a full-time faculty member available around the department and for office hours can be desirable. As a faculty member teaching it, it’s also a joy to chat with students about research in the hallway; teaching is fun, but talking about research and innovations in your field is exciting.

Teaching Research Methods and Statistics as a Faculty Member

There are other reasons that methods and statistics can be great to teach as a faculty member. First, you can bring in or read examples from any subdiscipline of psychology that you desire. Suppose you want to discuss the psychology of gender, organizational behavior, action research, the “classic” experiments, or even watch

Mythbusters (Burkley & Burkley, 2009). In that case, it's all immediately relevant without much of a stretch. For several years, my lab conducted a series of food-related cognition studies, and many of the examples I used in class were related to this area. The students got to eat a bunch of free candy, and we were able to test very controversial hypotheses about whether the black jellybeans are the worst (they are) – win–win.

You can also discuss research relevant to students' lives, whether within or outside of your field. At least within psychology, students tend to appreciate examples from helping professions on things like anxiety, depression, stigma, and socio-economic status (Sizemore & Lewandowski, 2011). It's also an opportunity to showcase the wide variety of applications and subdisciplines in the field. Students at smaller institutions with fewer courses may never get the chance to take classes in subdisciplines, yet they can be exposed to this research and see more career pathways than they previously imagined.

The topics related to research or statistics themselves can be exciting and relevant across a range of subdisciplines. For example, conversations related to research ethics can lead to some fascinating questions for students to address. How do you make sure that data are confidential and anonymous? Why is this important? How could a lack of confidentiality or anonymity negatively impact a person's life, and how could that affect the quality of the data? How has the scientific enterprise failed to earn the trust of people from different backgrounds or groups? For ethics-related issues, the reader might consult Chapter 2 in this volume.

Teaching research methods and statistics can increase your understanding of and familiarity with emerging issues and techniques across a range of subdisciplines. Research methods and statistics are not free of controversy, particularly since the early 2000s. This includes various debates and movements in the field, such as the replication crisis, the fact that most participants are from WEIRD countries (Western, educated, industrialized, rich, and democratic; Henrich et al., 2010), the racist beliefs of many statistical pioneers, and how they used statistics in service of these beliefs (e.g., Langkjær-Bain, 2019). Students can be drawn in and shown that research and statistics are not static sets of concepts but are, instead, part of a constantly evolving domain that informs and impacts all content areas in the field.

Unlike many other courses, a portion of the content does not need to change as regularly. I have some degree of faith that I will be able to use the same foundation for lectures on counterbalancing and *t*-tests for years to come. In contrast, in the other courses I primarily teach – Cognitive Psychology and Introduction to Psychology – the knowledge base changes so much from year to year that it can be challenging to keep up. Additionally, while content areas will change, the skills needed to evaluate claims and think about the field critically will not (or will at a significantly slower pace). It can be discouraging to think about the fact that so many things in our textbooks may be out of date within a decade or two, with us completely unable to predict what those things are. It is refreshing to teach a course that will remain broadly relevant for students decades later.

You might still be asking yourself if you should teach research methods and statistics. I would argue that you should. First, they are essential courses in many majors. Second, if you have a PhD in social or behavioral science, you likely have some research training and are qualified to teach it. Third, it can help bolster your research program by providing ideas for your work, bringing students into that work, or recruiting students to work in your lab. If you are at a primarily teaching institution, it can offer you the sort of close mentorship with students that is very rewarding. If you are at a large research institution, you can think more deeply about your research area. I often joke with students that I usually force myself to learn new things by agreeing to teach a course on it; research methods have that benefit for any scholar. When you talk about research, you are forced to think about your own research.

Recommendation 2: Cast a Wide Net for Inspiration

If you are teaching a research methods or statistics course for the first time, I encourage you to steal wheels rather than reinvent them. It is an open secret in academia that we all stand on each other's shoulders when it comes to course preparation; we give and receive course materials and ideas fairly freely (or at least, the people that I know do this). Sometimes, when you are teaching a very specialized course, this is difficult, but most of the time it's as easy as a google search. I do not mean you should copy and paste someone's syllabus as your own; I *do* mean you should look around for inspiration. Other instructors likely have the same learning outcomes as you and have many ways of meeting those learning outcomes.

I recommend starting with Society for the Teaching of Psychology (STP)'s project syllabus (Society for the Teaching of Psychology, n.d), which contains peer-reviewed syllabi in a range of topics. At this count, there are 22 syllabi listed in the Research Methods category, 14 under Statistics, and many more in other categories that may give you inspiration for ways to meet learning outcomes in your course. Some instructors, such as Dr. Morton Ann Gernsbacher from the University of Wisconsin–Madison, have made their entire courses available online (see https://online225.psych.wisc.edu for Dr. Gernsbacher's course) while others have posted their syllabi in public repositories (e.g., the Open Science Framework [OSF]; https://osf.io/vkhbt). Some scientists, like Dr. Jess Hartnett, comb the Internet to find fun things for us to use in our courses and post them on their blogs; Dr. Hartnett authors the site "Not Awful and Boring" (http://notawfulandboring.blogspot.com), which contains many hilarious and thought-provoking links to miscellaneous statistics and research examples – and lots of memes.

Recommendation 3: Make It an Opportunity for Justice, Diversity, Equity, and Inclusion

Teaching a course in research methods presents an excellent opportunity to discuss the ethical implications of research and how research might impact (or be impacted by) different communities; it may even create learning for students beyond the research methods material. For example, Yoder et al. (2016) were able to reduce students' sexist beliefs by assigning readings and activities throughout the semester related to sexism. The possibilities in your course could be endless. You can read journal articles in class that address the lived experiences of members of different communities, select readings by scholars outside of the global North, or select literature and research questions that impact women, families, communities of color, or vulnerable populations. You can use examples that have direct relevance to vulnerable populations. Finally, you can talk through how our identities inform the types of research questions we ask.

There are endless opportunities, in research methods and statistics courses, to introduce students to complex information that challenges the status quo. It may be particularly important for students to encounter this information in a course like research methods, so they can see that these issues are infused into the fabric of our field as well as how scholars can seek to dismantle oppressive systems with their work.

Recommendation 4: Include a Project

Students benefit from doing research projects in their research-related courses (Brownell et al., 2015). The finer details of which parts are or aren't helpful are less clear; you could absolutely use as an excuse to include the parts *you* think are important and remove the others. There is no "perfect project." Think about what skills you want students to learn, and make sure they practice these. If you don't think students will ever need to submit an IRB again in the future, maybe they don't need to do that part. However, I would recommend including some sort of data collection, some sort of hypothesis formation component (even if this isn't directly tested), some sort of writing, some sort of presenting, and some sort of collaboration. I also personally believe that students benefit from experience writing, revising survey questions, and seeing how other people read and respond to surveys. It is likely that, if students do go on to do some sort of research in the future, this may involve survey research – so this sort of experience would be helpful for them.

Most research courses require projects; in a sample of 62 research methods instructors, Gurung & Stoa (2020) found that 82% required some research project. The project itself can be a lot for an instructor to supervise, especially if the course is taught as a stand-alone, three-credit course in a department with limited teaching-assistant support. If students formulate and test their own individual hypotheses, the instructor may not know enough about the topics to provide good feedback. The instructor will need to read, evaluate, and provide feedback on each individual project; this may also include IRB submissions if the students ever wish to present the research beyond the walls of the classroom. Your department may put constraints on what types of projects can and should be done, such as original individual studies or group projects. Group projects may reduce the load related to feedback and supervision, but it also introduces a host of other issues, such as the work required to appropriately scaffold and support group work (Lou et al., 2000).

There are also issues that may seem less obvious as one begins teaching research methods – related to power, sampling bias, and methodological rigor. I will address these challenges in the next section and provide some possible solutions. These solutions are not comprehensive; no course has taxed my creativity like my research-related courses. For me, that's part of the fun – thinking of new ways to address old problems. For you, it might not be as fun, and that is when I recommend searching through the STP's project syllabus (Society for the Teaching of Psychology, n.d.) to get ideas. Do not reinvent the wheel. Someone, somewhere, has thought of the solution. If not, consider posting a question to the STP Facebook group (www.facebook.com/STP), Listserv (https://teachpsych.org/page-1862916), or even Twitter.

Recommendation 5: Prioritize Active Learning

Active-learning methods, as opposed to lecture-based teaching methods, enhance student learning (Freeman et al., 2014) and are effective in increasing student knowledge and confidence in research methods courses (Allen & Baughman, 2016; LaCosse et al., 2017). Active learning can include a variety of teaching methods, including polling the class, having students generate examples or responses, discussion, group work, in-class data collection and analysis, or conducting a study from beginning to end. There are many good resources available for incorporating active learning into research methods (e.g., Dawson, 2016; Stowell & Addison, 2017). However, transitioning from lecture delivery to active-learning delivery is no small task. If you are just starting out as an instructor, you might consider starting from an active-learning approach; if you have been teaching for a while, you could consider slowly replacing your lecture materials with demonstrations and activities.

Active-learning techniques are considered best practice in pedagogy (O'Neill & McMahon, 2005) and contribute to higher-level thinking outcomes (Richmond & Hagan, 2011). In research-related courses, active learning can be incorporated in many ways (e.g., short demonstrations, designing a survey together). Even short workshops where students collect real data and discuss it together have an advantage over looking at "canned" data – at least for the students' knowledge and confidence in the topics (Allen & Baughman, 2016). There is consensus in the literature that research projects confer advantages over other pedagogical approaches when it comes to learning outcomes (e.g., LaCosse et al., 2017).

Recommendation 6: Make It an Opportunity for Authentic Research

Research is a high-impact practice that helps with retention and engagement for all students (Kuh, 2008) but may be particularly helpful for first-generation students or students from marginalized or minoritized groups (Olson-McBride et al., 2016). Additionally, research experience is often necessary and always helpful for graduate-school admissions across a range of different disciplines (Landrum & Clark, 2005). However, not all students are aware of research opportunities outside of the classroom and, if they are, they may not feel comfortable approaching an instructor to ask if they can do research with them. They may also be unable to commit to extracurricular research opportunities because of family, work, transportation, or other obligations. Therefore, one way to "level the playing field" for these students may be to allow them to engage in authentic research experiences in their field (Grahe, 2017).

There are many ways to engage students in authentic research in the classroom, such as through service-learning experiences with community partners, by limiting their research projects to an area related to your own scholarship, or by having particularly industrious students who are up to the challenge of contributing original research to the field. I won't cover all of these, but I will say you should not be afraid of being creative with the project! Instead of taking a comprehensive approach, this section will focus on one type of project – the multisite collaboration (with an emphasis on replication projects). I focus on these projects for several reasons: (1) I know a lot about them, (2) I believe in them, and (3) I believe they are good for both students and instructors without increasing the workload associated with projects.

Multisite Collaborations

In 2012, Grahe and colleagues put out a call for undergraduate student research projects to contribute to the field. Large-scale collaborations, such as the Reproducibility Project (Open Science Collaboration, 2012), were already under way at the time; these collaborations specifically called for coordinated or crowd-sourced replication of published studies. Grahe et al. (2012) argued that big questions in psychology – not just replication studies, but novel research – could be coordinated so that *undergraduate students* would collect and analyze data related to those big questions as part of their coursework. The rationale was that the potential power of coordinated research efforts could contribute to the field quickly but could also meet the existing learning outcomes of the course without shifting instructional tools very much. Instead of students picking their own studies, groups formulate and test extension hypotheses on the research questions they contribute to. If anything, this sort of model might free up time for instructors who now only need to become mini-experts in one or maybe a handful of research questions.

There is also a need for replication research in the field, as we learned from the Reproducibility Project, and this sort of model would put some of the effort for replication in the hands of students. Although many researchers agree that replication is important, few engage in replication work because of the lack of incentives (Nosek et al., 2012); direct replications are important but difficult to publish (Schmidt, 2009). Because students typically do not need to get published to advance in their careers, replication research presents less risk for undergraduate or master's students than for doctoral students or faculty. Ironically, however, participation in coordinated efforts like the Collaborative Replications and Education Project (CREP; https://osf.io/wfc6u) might be *more* likely to result in students being published (Wagge et al., 2019), and this could be very useful for students who do want to advance in academia.

Here, I'll discuss CREP along with several other student-specific models. Other more general collaborative models also exist, such as the Psychological Science Accelerator (PSA) (Moshontz et al., 2018). The PSA is very student-friendly and warmly welcomes scientists from all sorts of institutions, ranks (including students), and continents. The focus in the PSA, though, is not on pedagogy but on decentralization and democratization of research priorities. Therefore, I encourage you to read more about the PSA, but I will primarily discuss initiatives with pedagogy as a primary focus.

The CREP

As mentioned, the CREP is a model for crowdsourcing student replication projects. As the current executive director for the organization, I feel I should declare this potential conflict of interest but also advertise the fact that all of our materials are publicly available; we encourage you to adapt these to your own needs, as you see fit. For example, you could download and adapt our manual for student projects for your own classes. You do not need to participate in CREP, although we definitely encourage it; if you find any of our processes or materials helpful, use them.

The CREP process starts with our organization picking studies that would be good to replicate. There are a variety of ways you could go about doing this, but we've chosen to look through a specific set of psychology journals from three years prior to the year we code (e.g., in 2016, we selected papers from 2013) and pull a list of the ten most-cited empirical papers since then. This gives us a list of around 100 studies; from that list, we start narrowing. We'll have raters (typically students) code the papers for feasibility – culling papers that involve things like functional MRI, clinical populations, or longitudinal research – and then our organization (mostly faculty members) will take this narrowed-down list of the most feasible studies and select the top papers (usually two or three) for CREP studies. If a paper has more than one study, we pick the focal study – usually the first study.

Once we've selected the studies, we prepare an OSF page for the study (e.g., https://osf.io/9sr2k; Sutherland et al., 2021). Then, we contact the corresponding author from the original study to let them know who we are and that we have chosen to replicate their paper. We ask them for any suggestions they may have for potential moderators and typically hear back quickly – with lots of helpful information. With the authors' permission, we will post these conversations on the OSF project page. Student teams then sign up to complete replications of that study, using our materials, and we have reviewers check their projects (including a procedural video) both before and after data collection. Throughout this process, students are engaging in some of the best practices of open and transparent science – preregistration of hypotheses on the OSF, open methods, and open data. They also learn why replication is important and why data from multiple sites is necessary to achieve statistical power goals. Further, they learn the methods employed by scholars in the field because they closely and carefully mimic, and then write about, those methods. Quintana (2021) argues that students should complete theses via replication efforts (or replication with extension) for the same reasons I argue for them to be completed in research courses – they benefit everyone without disadvantaging anyone.

I have employed CREP projects in my courses for several years now, and typically I will decide which project we will replicate and then allow students to add extension hypotheses in groups. Then, everyone writes a very similar paper with a slight difference in a section on their extension hypotheses. Students who complete projects as well as the CREP requirements are then invited to contribute to the manuscript when enough data are collected for a pooled analysis.

As mentioned, this model can be adapted to suit a variety of needs, which may or may not include replication. An instructor could take our step-by-step procedures and adapt them for original research questions in their class only; a group of instructors could also take our procedures and adapt them for a specific research question – there are a lot of possibilities. Teaching is hard – if you can save some time and energy *and* give your students a great experience meeting important learning outcomes without having to create everything yourself, do it! It's important to have time for things like video games and whatever it is other people do for self-care.

Other Multisite Projects

Table 36.1 gives a list of multisite projects, including several projects that are specific to students. Some of these projects are ongoing (e.g., Psi Chi's Network for International Collaborative Exchange), while others have finished but can provide a model for research in the future.

Single-Site Replication Efforts

Many researchers have published single-site replication efforts – either well-powered individual studies or combinations of multiple studies by multiple students. Hawkins et al. (2018) report the results from replications that were conducted in a graduate course; students selected recent papers from *Psychological Science*, preregistered their hypotheses, and conducted replication work. Along with Frank and Saxe (2012), they argue that replications need to be done and that students are in a unique position to do this work and to benefit from the experience.

Examples of single-site published direct replications that involve students abound in the literature. For example, Burns et al. (2019) attempted to replicate Adam and

Table 36.1 *Examples of multisite projects for researchers and students*

Type of project	Name of project
Student-specific	The Hagen Cumulative Science Project (Jekel et al., 2020) The GW4 Undergraduate Psychology Consortium (Button et al., 2020) Psi Chi's Network for International Collaborative Exchange (Cuccolo, 2019) EAMMI, EAMMI2, and EAMMI3 (e.g., Reifman & Grahe, 2016)
Not student-specific	Many Labs (Klein et al., 2014) Many Babies (Frank et al., 2017) The Psychological Science Accelerator (Moshontz et al., 2018)

Galinksy's (2012) work demonstrating that wearing a doctor's lab coat could improve performance on the Stroop task. Impressively, researcher Gilad Feldman from Hong Kong University has implemented a system for teaching and mentoring research entirely through replication studies (see https://mgto.org/pre-registered-replications). Many of these studies have resulted in publications (e.g., Chen et al., 2021).

Some instructors believe that having students formulate and test their own hypotheses is a priority, and that is not misguided; however, given the choice between having a student formulate and test a hypothesis that may not address a gap in the literature (and almost certainly will not be published or presented outside of the class) and having a student engage in research that they didn't conceptualize but is genuine scholarship that advances the field, I would choose the latter every time. Replication research conducted through the CREP, for example, offers students several opportunities for publication and citation (see Wagge et al., 2019) – students who conducted CREP research as part of their coursework are eligible to participate in the authorship process once the multisite study is complete. Additionally, engaging in authentic research allows research methods instructors to speak more thoroughly to the student's skills in letters of recommendation. Regardless of whether a student wants to attend graduate school, however, they deserve an opportunity to engage in authentic scholarship as part of their regular coursework. This also ties into the previous recommendation – what better way to dismantle inequity than to provide all of your students with opportunities that are often only afforded to a few?

Challenges in Teaching Research Methods and Statistics

Challenge 1: Enthusiasm

Perhaps the biggest challenge – at least from my perspective – to teaching these courses is the lack of enthusiasm from students entering the course. One potential solution is to take a "marketing" approach to get students enthusiastic about the course *and* the concept. This approach would need to involve the entire department, including advisors, instructors, and administration. However, faculty have limits to their enthusiasm; these are not popular courses to teach, as mentioned at the beginning of this chapter. The research methods instructor should be an ambassador for the topic to the rest of the program and the students. For example, I have learned over time not to make jokes about how students are unenthusiastic about these courses. Why wouldn't they be? The word "research" imparts something difficult – even if it is rewarding. Even students who enjoy doing difficult things might enter such a course with bated breath.

There are small ways that instructors could get their students enthusiastic about research methods. First, highlight the skills that will help students with employability – data analysis, critical thinking, problem solving. Second, highlight the things that college students enjoy – real-life examples and applications of research, philosophical discussions (e.g., the ethical questions raised earlier), creativity (the kind required to operationalize experimental manipulations). Third, highlight the characteristics of the research methods courses in your department when you are talking to students. Fourth, demonstrate enthusiasm about your research, scholarship, or experiences with research in the past. Model a positive attitude toward reading journal articles, attending conferences, and keeping up with the literature in the field; model a positive attitude toward research and statistics, in general. It will become a greater part of the culture in your department whether you regularly engage in those things or not.

Challenge 2: Course Content

Another big challenge is simply that there is so much to accomplish. For example, in the research methods courses I have taught, I have wanted students to leave the course with proficiency in an array of different skills and so at some point I have wanted to include all of the following: a research project, a literature review, SPSS (Statistical Package for the Social Sciences IBM package) activities and reporting, ethics training, plagiarism training, experience forming hypotheses (if not part of the research project), presenting work through oral presentations or posters, and reading journal articles. This is all on top of the typical research methods content that they need.

This is quite a bit, although I have found ways to make this work by having assignments meet multiple goals *and* by paring down assignments to impart the skills I want students to leave with. For example, they might complete a direct replication project (for which I have already written the IRB). Then, they write an empirical report – including a short literature review – and propose extension hypotheses or conceptual replications in a final presentation. Along the way, they also complete SPSS assignments using our own data from the projects. I assign specific readings that they need to incorporate into their literature review; this gives them experience reading journal articles *and* has the added benefit of allowing me to readily critique each student's paper due to my familiarity with the topic. I have now included all the assignments that I set out to incorporate in that original list, and they all connect to that one research project. This might look different to you, but I'd encourage you to think about how these things can overlap.

Let's focus just on research methods (although much of what is discussed here could be generalized to statistics). A lot gets accomplished in a typical research methods semester, and it has so many of the things that students often find intimidating – software, writing, math, reading. There may be some variation in how topics in research methods are covered. Still, there does at least appear to be some sort of "core" curriculum (Gurung & Stoa, 2020), with most instructors agreeing on the concepts (e.g., ethics) that should be covered regardless of areas of expertise (e.g., psychology, sociology) or level of the student (e.g., graduate student, full professor). There are also many important topics that don't seem to be covered well in most research methods courses or textbooks; these include qualitative research, mixed methods research, and cross-cultural issues. For a more thorough treatment of these topics, you could consult many of the chapters in this handbook.

Ethical issues are fundamental to discuss within research methods courses. I would recommend infusing them throughout the course rather than as a stand-alone unit at the beginning of the class. Ethical discussions within a research framework should go beyond the checklist – what we need to do a study. I would argue that, without research ethics discussions, we cannot study human behavior or mental processes in a meaningful way. Knowing that researchers have harmed some communities, for example, may help students understand why some participants are hesitant to volunteer for research or respond to requests for research participation. We tend to talk about these things during the research ethics chapter, but other topics are rarely addressed in standard research methods textbooks and are also important; not only are these topics important, but students (at least appear to) enjoy talking and thinking about them in class.

At least one textbook focuses solely on ethical issues in psychological research (Corts & Tatum, 2019). Still, many general research methods texts fail to acknowledge the ethical implications of the decisions that researchers (and others) need to make at every step of the process. Deciding on participants, designs, publication outlets, research team members, statistical analyses, what to report, and what not to report can all have ethical implications and are essential to discuss with students.

Challenge 3: Workload

I would also encourage you to be completely unapologetic in overlapping the needs of your own research program with the class. Instead of replication work, for example, you might have students learn about your research and collect data for your next study. They might also propose “next steps” or additional hypotheses in a final presentation. Here are some potential advantages to doing this: (1) students get to learn about a research topic from an expert that is teaching their research methods course, (2) students still get to “make it their own” by proposing an extension, (3) you learn exactly how interpretable your research is by novices in the field – this can help you in communicating that work to the general public, (4) you might get some very creative next-step research ideas from students, and (5) related to the last point, you might end up inviting students to do research with you and publish a paper with data you collected in the class.

I’d also like to throw in one more advantage – grading papers and giving feedback will be easier and likely more meaningful when you are familiar with the research area they are discussing. I would probably need eight dozen limbs to be able to count on my fingers the number of times I’ve read literature reviews on attachment and have thought to myself that I wish I knew the literature better. For papers out of my area of expertise, I may be able to identify general flaws in logic and poor synthesis or rationale for hypotheses, but I cannot often tell whether they are summarizing the research accurately. If I am reading papers whose background literature I’m more familiar with, I can give more meaningful feedback on whether students are interpreting the literature correctly or providing good insights about the field. This is an advantage for the students and myself.

Challenge 4: Making Projects Matter

Perhaps one of the biggest decisions to make as an instructor, working on projects with students, is whether those data will be presented outside of class. I would argue that this should generally be a goal. Most fields already have a significant "file drawer" problem – unknown but likely vast amounts of data have been neither published nor made publicly available in any format (perhaps most typically because of the "aversion to the null"; Heene & Ferguson, 2017). It is my estimation that the file drawer full of undergraduate research project findings have a mix of effects and have been mostly filed and drawered because there have been no clear paths to or incentives for making the data public; also, students and instructors often move on after the semester is over (me included). However, if instructors adopted the use of public repositories (e.g., the OSF) and incorporated posting anonymized data into students' final grades for projects, then this could be less of a problem. Millions of data points are gathered by students each semester, and very few are presented outside of class. Perlman & McCann (2005) estimate this is less than 10%, and I would gather that most of that work is presented in department, university, or local conferences. This sort of distribution is also very limited and tends to not really be "public" – future researchers cannot typically access the presentation or the materials or data.

One might argue that, given the lack of methodological rigor present in many undergraduate research projects (Wagge et al., 2022), we could be grateful that these data have not typically entered the public domain. I would argue that even these data could be valuable to researchers in the future, who may have questions about how novices design studies or how participants might interpret poorly worded questions. We cannot anticipate these things, but we can try to guarantee that the work of our students and our participants is honored by making the results of research available to others who may want access to it in the future.

Sometimes, it is not possible. For example, if you collect data that cannot reasonably be deidentified, you should not post raw data online. You can, however, post summary data and effect sizes alongside your methods, which would be of great use to meta-analytic researchers. You can also simulate data from the *faux* package in R (DeBruine, 2020); you would just want to be clear in your description of the data that they have been simulated. This can offer students a way to showcase their data analysis skills without worrying about identification of participants.

Challenge 5: Slow IRB Processes

There is one more major hurdle to consider here – the IRB process. This is required for students at most institutions in the United States prior to presenting data outside of class. After considering the obstacles listed above, if you believe that data from student projects in your class are very unlikely to be helpful to any future researchers, your students may not want to present their research somewhere (including on-campus events), *and* you have a cumbersome IRB process at your institution, you may decide against garnering IRB approval. There may be other reasons I am overlooking as well, but in my work with instructors who supervise projects (especially those at smaller institutions), the length of time it takes for an IRB to respond is a primary factor in *not* seeking approval.

Even if your IRB is slow, this is an obstacle that may be overcome by teaching statistics and methods as an integrated, two-semester sequence; this potentially buys up to 32 weeks (two 16-week semesters) of class time with perhaps 4–6 weeks in between semesters, allowing more of a cushion for the IRB. At my institution, we teach a lot of transfer students who may have the first semester but not the second or who take either the first or second course in our evening “adult learner” abbreviated program. That means that we’ll lose students after the first semester who have proposed their research, and we will gain some students the second semester who will need to be incorporated into existing groups. One potential solution to this problem is to submit IRB applications prior to the start of the semester, but that involves either (a) getting blanket approval for any topic selected by your students – this can be tricky to navigate, or (b) selecting the project(s) they will work on prior to them even starting the class.

The latter option may still provide some degree of choice for the student. For example, you may select three or four different studies students can choose from to complete or replicate and submit IRB applications for them. Going a step further, if your IRB *modification* process is much quicker than the approval process, you could even have students submit extension hypotheses of their own design. This is one possible model for the CREP in the classroom, which I will discuss below. If you’re at a small institution, it might also be worth having a conversation with your IRB to see if they could provide approval or feedback within a certain range of time if the materials are submitted by a specific date. At a smaller teaching institution, it may be helpful to frame this request in terms of research being a high-impact practice for students. At the very least, if there is a well-established long turnaround time for IRB reviews at your institution, then work with your chair, dean, or provost to try to find a solution to the problem before abandoning hope.

If you have abandoned hope, though, another solution is to have students do archival work, content analysis, or form hypotheses about existing data. Students could, for example, examine public blog posts, tweets, or other records for specific characteristics. They could code journal articles related to some characteristic, such as how often a specific concept is mentioned or how many times the author cites themselves. Finally, there have been large data sets produced for this purpose; this includes the project Emerging Adulthood Measured at Multiple Institutions (EAMMI) (Reifman & Grahe, 2016) along with its sequels, EAMMI2 and the upcoming EAMMI3. EAMMI and EAMMI2 have both provided rich data sets that characterize the developmental period of emerging adulthood (age 18 to 29) and is relevant for many students in undergraduate courses. Exploration of these data sets has resulted in publications for both students and faculty (e.g., Cuccolo et al., 2021; Long & Chalk, 2020; Skulborstad & Hermann, 2016).

Challenge 6: Doing Rigorous Projects

Other potential problems include issues of power, sampling bias, and methodological rigor. Student projects rarely have enough power to form appropriate statistical conclusions (Wagge et al., 2022). Still, students often examine data and report results without acknowledging potential Type I or Type II errors due to low sample size. When students can collect a lot of data, it is often from their contacts and, therefore, could have issues with sampling bias (although likely not anything more serious than the sampling bias we see in normal university participant pools). Finally, student projects often feature ad hoc methods, such as surveys explicitly written for their project that have not been subject to pilot testing or validation techniques. The psychometric properties of student projects may be questionable – yet reasonable for novices in the field – and the instructor may not be able to identify and discuss potential issues related to every student's project. This may perpetuate the idea that it is okay to create and distribute ad hoc measures (Flake & Fried, 2020). Given some of the challenges that face the field because of poor psychometric properties of ad hoc measures, I would guard against this.

Assuming that funds are not available to compensate participants for in-class projects, there are a few other possible ways to address the obstacles related to lower power. First, if available, use a participant pool; this is not an option at all institutions, however. Another option would be to participate in a multisite collaboration – either one you organize yourself or an established collaboration that you can join. Multisite collaboration models were discussed above and I recommend referring back to this section for ideas if you are interested in either joining an existing project or creating your own.

For obstacles related to methodological rigor, you can take several approaches. You might create a scale together as a class, or you might ask students to address psychometric properties as part of their paper. The benefit of creating a scale together as a class is that it gives you an opportunity to really dig into survey creation. This is a skill that many students will use again as part of their work, or perhaps even with polls they put on social media, so knowing how to write items for and analyze results from a good survey could be impactful. Here, I look back on my experience as a political science major prior to switching to psychology as an undergraduate – taking methods and statistics courses in each prior to graduation. The political science courses emphasized survey design and analysis while the psychology courses emphasized experimental design and analysis; I am grateful for having both experiences. Writing good surveys is hard and takes work, and students can benefit from active-learning experiences here.

Another solution to the rigor problem is to use established scales or operationalizations that have good validity and reliability within the population being tested. I have found that this is often difficult for students and requires close mentorship when selecting scales. A third option, which I have tried, is to have an existing set of scales that students can choose from to form hypotheses. One benefit of this model is that it was easy to get blanket IRB approval at my institution for surveys containing any number of these scales along with demographic information. Using this approach, you could gather a potentially large number of different measures, and then students could use theory and the literature to predict the relationships between different scales. Using this approach once in a summer course for graduate students, a small number of students each selected scales and a survey was then compiled with each. The survey took participants around 30 minutes to complete, but because every student was collecting data, we were able to gather data from close to 200 participants. The students then used the research they conducted to formulate research proposals for next steps. Therefore, despite a large part of their work being "canned" or predictated, they had many opportunities to make choices and be creative in their work.

Conclusion

Research methods and statistics can feel like more challenging courses to teach than they really are. I encourage you to try teaching them at least a few times and make them your own. You can do so many things in these courses that will make them meaningful to both you and your students. As a result, you may be pleasantly surprised with your experiences once you're done.

References

Adam, H. & Galinsky, A. D. (2012). Enclothed cognition. *Journal of Experimental Social Psychology*, *48*(4), 918–925. https://doi.org/10.1016/j.jesp.2012.02.008

Allen, P. J. & Baughman, F. D. (2016). Active learning in research methods classes is associated with higher knowledge and confidence, though not evaluations or satisfaction. *Frontiers in Psychology*, 7, March 1. https://doi.org/10.3389/fpsyg.2016.00279

Brownell, S. E., Hekmat-Scafe, D. S., Singla, V., et al. (2015). A high-enrollment course-based undergraduate research experience improves student conceptions of scientific thinking and ability to interpret data. *CBE – Life Sciences Education*, *14*(2), ar21. https://doi.org/10.1187/cbe.14-05-0092

Burkley, E. & Burkley, M. (2009). Mythbusters: A tool for teaching research methods in psychology. *Teaching of Psychology*, *36*(3), 179–184. https://doi.org/10.1080/00986280902739586

Burns, D. M., Fox, E. L., Greenstein, M., Olbright, G., & Montgomery, D. (2019). An old task in new clothes: A preregistered direct replication attempt of enclothed cognition effects on Stroop performance. *Journal of Experimental Social Psychology*, *83*, 150–156. https://doi.org/10.1016/j.jesp.2018.10.001

Button, K. S., Chambers, C. D., Lawrence, N., & Munafò, M. R. (2020). Grassroots training for reproducible science: A consortium-based approach to the empirical dissertation. *Psychology Learning & Teaching*, *19*(1), 77–90. https://doi.org/10.1177/1475725719857659

Chen, J., Hui, L. S., Yu, T., et al. (2021). Foregone opportunities and choosing not to act: Replications of inaction inertia effect. *Social Psychological and Personality Science*, *12*(3), 333–345. https://doi.org/10.1177/1948550619900570

Corts, D. P. & Tatum, H. E. (2019). *Ethics in Psychological Research: A Practical Guide for the Student Scientist*. SAGE Publications.

Cuccolo, K. (2019). Engaging in cross-cultural research with Psi Chi's Network for International Collaborative Exchange (NICE). *Eye on Psi Chi*, *23*(4), 46–47. https://doi.org/10.24839/2164-9812.Eye23.4.46

Cuccolo, K., Irgens, M. S., Zlokovich, M. S., Grahe, J., & Edlund, J. E. (2021). What crowdsourcing can offer to cross-cultural psychological science. *Cross-Cultural Research*, *55*(1), 3-28.

Dawson, C. (2016). *Activities for Teaching Research Methods*. SAGE Publications.

DeBruine, L. (2020). faux: Simulation for factorial designs (0.0.1.2) [Computer software]. https://doi.org/10.5281/ZENODO.2669586

Flake, J. K. & Fried, E. I. (2020). Measurement Schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Frank, M. C., Bergelson, E., Bergmann, C., et al. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy: The Official Journal of the International Society on Infant Studies*, *22*(4), 421–435. https://doi.org/10.1111/infa.12182

Frank, M. C. & Saxe, R. (2012). Teaching replication. *Perspectives on Psychological Science*, *7*(6), 600–604. https://doi.org/10.1177/1745691612460686

Freeman, S., Eddy, S. L., McDonough, M., et al. (2014). Active learning increases student performance in science, engineering, and mathematics. *PNAS*, *111*(23), 8410-8415. www.pnas.org/cgi/doi/10.1073/pnas.1319030111

Friedrich, J., Buday, E., & Kerr, D. (2000). Statistical training in psychology: A national survey and commentary on undergraduate programs. *Teaching of Psychology*, *27*(4), 248–257. https://doi.org/10.1207/S15328023TOP2704_02

Grahe, J. E. (2017). Authentic research projects benefit students, their instructors, and science. In R. Obeid, A. Schwartz, C. Shane-Simpson, & P. J. Brooks (eds.), *How We Teach Now: The GSTA Guide to Student-Centered Teaching* (pp. 351–367). Society for the Teaching of Psychology. Available at: http://teachpsych.org/ebooks/howweteachnow.

Grahe, J. E., Reifman, A., Hermann, A. D., et al. (2012). Harnessing the undiscovered resource of student research projects. *Perspectives on Psychological Science*, *7*(6), 605–607. http://doi.org/10.1177/1745691612459057

Gurung, R. A. R. & Stoa, R. (2020). A national survey of teaching and learning research methods: Important concepts and faculty and student perspectives. *Teaching of Psychology*, *47*(2), 111–120. https://doi.org/10.1177/0098628320901374

Hawkins, R. X. D., Smith, E. N., Au, C., et al. (2018). Improving the replicability of psychological science through pedagogy. *Advances in Methods and Practices in Psychological Science*, *1*(1), 7–18. https://doi.org/10.1177/2515245917740427

Heene M. & Ferguson C. J. (2017). Psychological science's aversion to the null, and why many of the things you think are true, aren't. In S. O. Lilienfeld & I. D. Waldman (eds.), *Psychological Science under Scrutiny: Recent Challenges and Proposed Solutions* (pp. 34–52). John Wiley & Sons.

Henrich, J., Heine, S. J., & Norenzahan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–135.

Jekel, M., Fiedler, S., Allstadt Torras, R., et al. (2020). How to teach open science principles in the undergraduate curriculum: The Hagen Cumulative Science Project. *Psychology Learning & Teaching*, *19*(1), 91–106. https://doi.org/10.1177/1475725719868149

Klein, R. A., Ratliff, K. A., Vianello, M., et al. (2014). Investigating variation in replicability: A “many labs” replication project. *Social Psychology*, *45*(3), 142–152. http://dx.doi.org/10.1027/1864-9335/a000178

Kuh, G. D. 2008. *High-Impact Educational Practices: What They Are, Who Has Access to Them, and Why They Matter*. Association of American Colleges and Universities.

LaCosse, J., Ainsworth, S. E., Shepherd, M. A., et al. (2017). An active-learning approach to fostering understanding of research methods in large classes. *Teaching of Psychology*, *44*(2), 117–123. https://doi.org/10.1177/0098628317692614

Landrum, R. E. & Clark, J. (2005). Graduate admissions criteria in psychology: An update. *Psychological Reports*, *97*(2), 481–484. https://doi.org/10.2466/pr0.97.2.481-484

Langkjær-Bain, R. (2019). The troubling legacy of Francis Galton. *Significance*, *16*(3), 16–21.

Long, O. & Chalk, H. M. (2020). Belonging and marital perception variances in emerging adults with differing disability identities. *Psi Chi Journal of Psychological Research*, *25*(1), 22–29. https://doi.org/10.24839/2325-7342.JN25.1.22

Lou, Y., Abrami, P. C., & Spence, J. C. (2000). Effects of within-class grouping on student achievement: An exploratory model. *The Journal of Educational Research*, *94*(2), 101–112. https://doi.org/10.1080/00220670009598748

Moshontz, H., Campbell, L., Ebersole, C. R., et al. (2018). The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*, *1*(4), 501–515. https://doi.org/10.1177/2515245918797607

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631. https://doi.org/10.1177/1745691612459058

O’Neill, G. & McMahon, T. (2005). Student-centred learning: What does it mean for students and lecturers? In B. O. Neill, S. Moore, & B. McMullin (eds.), *Emerging Issues in the Practice of University Learning and Teaching* (pp. 27–36). All Ireland Society for Higher Education.

Olson-McBride, L., Hassemer, H., & Hoepner, J. (2016). Broadening participation: Engaging academically at-risk freshmen in undergraduate research. *Council on Undergraduate Research Quarterly*, *37*(1), 4–10.

Open Science Collaboration (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*(6), 657–660. https://doi.org/10.1177/1745691612462588

Perlman, B. & McCann, L. I. (2005). Undergraduate research experiences in psychology: A national study of courses and curricula. *Teaching of Psychology*, *32*(1), 5–14. https://doi.org/10.1207/s15328023top3201_2

Quintana, D. S. (2021). Replication studies for undergraduate theses to improve science and education. *Nature Human Behaviour*, *5*, 1117–1118. https://doi.org/10.1038/s41562-021-01192-8

Rajecki, D. W., Appleby, D., Williams, C. C., Johnson, K., & Jeschke, M. P. (2005). Statistics can wait: Career plans activity and course preferences of American psychology undergraduates. *Psychology Learning & Teaching*, *4*(2), 83–89. https://doi.org/10.2304/plat.2004.4.2.83

Reifman, A. & Grahe, J. E. (2016). Introduction to the special issue of emerging adulthood. *Emerging Adulthood*, *4*(3), 135–141. https://doi.org/10.1177/2167696815588022

Richmond, A. S. & Hagan, L. K. (2011). promoting higher level thinking in psychology: Is active learning the answer? *Teaching of Psychology*, *38*(2), 102–105. https://doi.org/10.1177/0098628311401581

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*(2), 90–100. https://doi.org/10.1037/a0015108

Sizemore, O. J. & Lewandowski, G. W. (2011). Lesson learned: Using clinical examples for teaching research methods. *Psychology Learning & Teaching*, *10*(1), 25–31. https://doi.org/10.2304/plat.2011.10.1.25

Skulborstad, H. M. & Hermann, A. D. (2016). Individual difference predictors of the experience of emerging adulthood. *Emerging Adulthood*, *4*(3), 168–175. https://doi.org/10.1177/2167696815579820

Society for the Teaching of Psychology (n.d.) Project syllabus description. Available at: http://teachpsych.org/otrp/syllabi/index.php.

Stowell, J. R. & Addison, W. E. (eds.). (2017). *Activities for Teaching Statistics and Research Methods: A Guide for Psychology Instructors*. American Psychological Association. https://doi.org/10.1037/0000024-000

Sutherland, C. L., Hildebrandt, L., Wagge, J. R., et al. (2021). Troy, Ford, McRae, Zarolia, & Mauss (2017). https://doi.org/10.17605/OSF.IO/9SR2K

Wagge, J. R., Brandt, M. J., Lazarevic, L. B., et al. (2019). Publishing research with undergraduate students via replication work: The Collaborative Replications and Education Project. *Frontiers in Psychology*, *10*, 247. https://doi.org/10.3389/fpsyg.2019.00247

Wagge, J. R., Hurst, M. A., Brandt, M. J., Lazarevic, L. B., Legate, N., & Grahe, J. E. (2022). Teaching research in principle and in practice: What do psychology instructors think of research projects in their Courses? Psychology Learning & Teaching, 0(0). https://doi.org/10.1177/14757257221101942

Yoder, J. D., Mills, A. S., & Raffa, E. R. (2016). An effective intervention in research methods that reduces psychology majors' sexist prejudices. *Teaching of Psychology*, *43*, 187–196. https://doi.org/10.1177/0098628316649314

37 Working Outside Academia

Kevin A. Byle, Jeffrey M. Cucina, Alexis B. Avery, and Hanna K. Pillion

Abstract

Working in applied settings presents unique challenges and complexities with respect to research. Researchers have often commented on the scientist–practitioner divide, but there is a lack of information about the specific challenges and constraints of doing applied work that may contribute to this divide. As a group of applied social and behavioral scientists, we discuss what individuals should know, understand, and expect regarding the work practitioners conduct in applied settings. We describe the challenges of applied work as they relate to some topics covered earlier in this volume and identify other unique aspects of applied work. We conclude by discussing how an individual can approach deciding whether applied work is a fit with one's interests.

Keywords: Practitioner, Organizations, Applied Work, Statistics, Methodology, Career

Introduction

Working outside academia can provide fulfilling opportunities and careers. Often, individuals contemplate working outside academia to do applied work. Sometimes, that occurs while completing an advanced degree, as students decide between an academic or applied career path. Other times, academics desire to make a career change or expand their research into applied settings. The purpose of this chapter is to delineate the common challenges of applied work as they relate to earlier sections in this volume, where applicable. Researchers have frequently commented on the scientist–practitioner divide (Aguinis & Pierce, 2008; Anderson, 2007; Cascio, 2008; Gelade, 2006; Hodgkinson, 2006; Rynes, 2007, 2012; Rynes et al. 2001), but there is a lack of information regarding the specific practical challenges of doing applied work – some of which may contribute to the difference between the "how it should be done" scientist perspective and the "how it is actually done" practitioner experience. As a group of practitioners, we describe, from an experiential perspective, what individuals should know, understand, and expect regarding the work conducted in applied settings.

Author note: The views expressed in this chapter are those of the authors and do not necessarily reflect the views of US Customs and Border Protection or the US Federal Government.

This chapter is beneficial for a variety of audiences. It may benefit those currently completing their advanced degree, who are deciding between academia and applied work. It also may benefit academics teaching future practitioners or contemplating working outside academia themselves. Individuals are often surprised or even discouraged when they conduct applied work for the first time, only to realize the differences between theory and practice. Internships, while certainly helpful in providing students a glimpse of what working in applied settings is like, typically lack the responsibilities of a full-time position and may not be lengthy enough to allow for a complete understanding of academic–applied differences. There are also many differences across applied jobs. A principal scientist or researcher for a testing firm may be heavily involved in research. A consultant may have responsibilities to network, develop sales leads, and sell products and services. A practitioner internal to an organization may have to run program operations and manage a team. Because of the differences between academia and applied settings, as they relate to work and the many responsibilities practitioners can have, we provide a much-needed resource to consult when considering applied work.

In this chapter, we describe the challenges of applied work, as they relate to topics covered earlier in this volume, and discuss other unique aspects of applied work that differ from academics. Specifically, we first describe the research process in applied settings. Next, we discuss the typical statistical techniques used in applied settings along with common challenges that practitioners experience regarding data. Third, we address the challenges experienced in applied work, and critical skills and duties practitioners can expect to perform above and beyond their academic training. We conclude by discussing how career paths, promotional opportunities, and work characteristics differ between academic and non-academic settings; we also discuss how individuals interested in applied work can best approach deciding whether it is a good fit.

The Applied Research Process

The scientific method provides a good framework for designing and conducting applied research. Cucina et al. (2014b, p. 357) reviewed textbooks and other references from several fields of science (e.g., chemistry, physics, psychology) and observed that the scientific method is often described as consisting of the following steps:

(1) Make an observation.
(2) Form a question.
(3) Write a hypothesis.
(4) Make a prediction (i.e., if the hypothesis is true, then the prediction will be true).
(5) Test hypothesis using experimentation or observation.
(6) If the test supports hypothesis, then make new tests for hypothesis. If the test does not support hypothesis, then revise or create new hypothesis.
(7) Repeat steps 1 through 6 many times. Only if a hypothesis is supported after many replications can it become a theory.

There are some divergences between the scientific method as used in academic and basic research versus how it is implemented in research conducted by practitioners. In academic research, the ultimate goals are to increase and share understanding of behavior and phenomena and to build a body of knowledge with generalizable conclusions. In applied research, the goal is to conduct research and implement interventions to improve very specific aspects of the organization. For example, an academic might be interested in studying the relationship between personality constructs and sales performance, with particular attention on whether relationships generalize across different types of sales jobs. In contrast, a practitioner might be tasked with studying the relationship between scores on a personality test and sales performance for a single occupation with the goal of improving selection for that specific occupation. In this instance, the research topics are similar but have a different scope, focus, and purpose.

Applied research often begins with an observation (i.e., step 1 of the scientific method) made by an organization's decision makers, leaders, or stakeholders. For instance, observations might be made that new hires lack critical competencies, low morale exists in the workforce, or that the demographics of the workforce do not match the populations it is serving. Observations such as this lead to applied research questions (e.g., What critical competencies do new hires lack? or Why is workforce morale low?) and speculations (i.e., hypotheses) about the answers to those questions – steps 2 and 3 of the scientific method.

It is often at this point that practitioners become involved and use their expertise and experience in research methodology and statistics to develop a study to test predictions about the research question – steps 4 and 5 of the scientific method. For example, in response to the question, What critical competencies do new hires lack?, a practitioner might design a job analysis study to identify the critical competencies for the identified positions. Hypotheses about which competencies are critical could be made and tested using a job analysis survey and linkage ratings – ratings of the link between the competencies and the duties or tasks for the position. Additional tests could be conducted to fulfill step 6 of the scientific method (e.g., a criterion-related validation study on the relationship between scores on measures for critical competencies and job performance).

The last step of the scientific method (step 7) involves repeating the earlier steps until enough hypothesis testing has been conducted to establish a scientific theory. This is the major difference between applied research conducted in non-academic settings and basic research conducted in academia. In some cases, an applied researcher may wish to pursue theory development (e.g., by repeatedly testing a hypothesis using participants from different settings and measures of the relevant constructs). For instance, validity generalization of tests of general mental ability predicting job performance was primarily established using data collected by the US Federal Government and practitioners working for it (Schmidt & Hunter, 2003). However, in many applied settings, those sponsoring the research may not have the resources, population, or desire to support theory development.

Lack of Design Control

Oftentimes, the increased ecological validity associated with applied research comes at the expense of reduced internal validity. Applied research often lacks the experimental and research design control that academic research affords. Indeed, many applied research studies are field studies and quasi-experiments; Cook and Campbell (1979) described a number of concerns related to these types of studies. For example, a practitioner might investigate the efficacy of an intervention using a study that lacks a control group, or there could be contamination with the control group (e.g., members of the intervention group interacting with those from the control group).

Many practitioners are involved in survey programs (e.g., annual job satisfaction surveys). After job satisfaction surveys, focus groups are often conducted with employees to understand low-scoring areas, action plans are developed, and interventions are implemented to address areas of concern. This process provides a good example of how a lack of experimental design and control makes it difficult, if not impossible, to make conclusions about intervention efficacy. Focus groups might find that specific leadership skills are lacking and, as a result, an employer delivers leadership training to address the deficit. If job satisfaction scores rise the next year the survey is administered, it is not possible to make inferences of causality with respect to the training and job satisfaction scores. As a practitioner, one cannot conclude whether the score increase was due to the intervention or other action planning activities, initiatives, influences, external events (e.g., the organization’s profit, the current job market), or even self-improvement on the known deficit.

Research Summarization

After research has been conducted, findings typically are communicated to others within (and sometimes outside of) an organization. There are typically three audiences for practitioners: other practitioners, non-technical stakeholders, and the research community.

Technical reports, manuals, and research notes are typically used for communicating research to other practitioners, and they serve two purposes. First, they provide other practitioners with complete documentation of a study. For example, a well-documented job analysis study can be subsequently used by other practitioners to develop structured interviews, written tests, training programs, performance appraisals, and other instruments. Second, they provide the necessary level of technical documentation needed if practices are challenged (e.g., a selection system). For instance, if there is a claim of disparate impact involving a selection instrument, copies of the job analyses, test manuals, and criterion-related validity studies might be provided to legal counsel as evidence. Employers are expected to maintain job analysis and test validation reports in accordance with the Uniform Guidelines on Employee Selection Procedures (29 CFR Part 1607); they can be required to make these reports available to federal agencies who use the Uniform Guidelines to enforce federal anti-discrimination laws. When legal challenges occur, the technical materials could be scrutinized by expert witnesses, attorneys, statisticians, and others. For example, the methodology of a test validation study and the need for conducting a job analysis was the central focus in the *Albemarle Paper Company* v. *Moody* (1975) case,[1] which was ultimately decided by the US Supreme Court (see Johnson, 1976).

The second audience is non-technical stakeholders – executives, managers, supervisors, employees, union officials, oversight entities (e.g., auditors), and owners (e.g., a single owner, shareholders, or the entire citizen population, in the case of government organizations). The exact format for this reporting largely depends on the recipient, their format preferences, the context, the organization's culture, and other factors. Examples of this are one- or two-page issue papers or fact sheets, a bullet-point format that can easily be skimmed, and PowerPoint presentations slides.

The third audience, the research community, is often optional, although some practitioners are expected or encouraged to publish or have a desire to share their findings with others. Sometimes, the results of a study may be presentable at a scientific meeting, publishable in a peer-reviewed journal, or turned into a book chapter – especially when the topic is novel. However, it is unlikely that some types of work (e.g., an ordinary job analysis, validation study, or test development effort) will be accepted for these outlets; case studies can be an exception.

Due to the various audiences, messaging often needs to be tailored to the recipient. Practitioners need to discern between what technical information is critical to include or leave out in a presentation. There is a delicate balance between presenting too much technical information (and risk confusing or alienating the audience) and presenting too little (and risk losing credibility). In our experience, most stakeholders understand percentages, means, and interpreting Likert-scale responses – although, the percent of positive ratings can ease interpretation, and some stakeholders understand methodology. Rather than assuming this is true for all audiences, we have found it helpful to try to ascertain the level of understanding of technical information of the audience beforehand (e.g., through a staff member who knows or is experienced with those attending a meeting). It can also be helpful to create presentations with general information but prepare backup slides or files that can be shown if technical questions are asked.

If stakeholders do ask technical questions, practitioners should be prepared to provide some interpretative information. The US Department of Labor's (2000, pp. 3–10) guidelines for interpreting criterion-related validities (e.g., "above .35 very beneficial, .21 – .35 likely to be useful, .11 – .20 depends on circumstances, and below .11 unlikely to be useful") are a good example. Essentially, whenever a statistic is presented, some brief and concise explanation of what it means should accompany it.

Data and Statistical Considerations

One benefit of working in non-academic settings is that they provide a source of rich information, and this allows practitioners the opportunity to conduct important applied research. However, there are some commonly experienced

[1] *Albemarle Paper Co*. v. *Moody*, 422 US 405 (1975).

constraints that practitioners encounter when doing research. We describe several ways in which working with data in applied settings may be different than academic settings.

Incomplete or Missing Data

A first issue that can be experienced when working in applied settings is that data may simply not exist for important variables you desire to use. While most organizations tend to be very data-oriented, the systems that are used to store and capture data are often not set up in a manner that a researcher or practitioner needs. When a system is set up for data storage, practitioners are often not consulted or involved in determining how the data should be captured. Other professionals (e.g., database administrators, human resource specialists, information technology specialists) who are involved in system setup often do not have expertise in measurement, know what variables are needed, or know how to capture variables in a manner a practitioner needs.

Other times, as we have experienced as practitioners, data do exist but lack completeness needed for proper research and data analysis. Examples of this are data that are omitted or data categories that are not captured but are necessary to conduct data analysis. For example, a structured interview may have several competencies on which job applicants are rated, but the data storage system housing the structured interview data might only contain whether the job applicant passed or failed. This becomes problematic when a researcher wants or needs to use the ratings for research purposes.

A second and related issue is that a data storage system may capture only some of the necessary categories needed for analysis. For instance, data tracking employee turnover and separations may indicate who has left the organization, but a limited number of turnover reason categories may be recorded. Did the employee leave because he or she was a poor performer, unsatisfied, or simply found a better job elsewhere? The limitations of the existing data categories may not provide a sufficient level of differentiation or information needed to answer questions such as this. Other common issues one can encounter are receiving data that are not cleaned or formatted correctly or have duplicate or conflicting information. Each of these add steps and makes work and research in applied settings more complicated and time-consuming.

Employee Identification

Another frequent issue that practitioners experience is employee identification. Organizations have methods of tracking employees on an individual level for various purposes, and this is typically accomplished by using a unique identifier (e.g., Social Security Number [SSN] or Employee Identification Number [EIN]). Unique identifiers help connect data stored in separate locations; however, there are times when data sets lack a consistent unique identifier, making it difficult to match data from multiple sources. Additionally, organizations are increasingly shifting away from using SSNs to protect sensitive identifiable information, particularly when tracking job applicants.

Because much of the data in organizations are linked to individual employees, it introduces the task of managing sensitive personally identifiable information (PII). PII is any information that permits the identity of an individual to be directly or indirectly inferred, including any other information that is linked or linkable to an individual. Sensitive PII is information that, if lost, compromised, or disclosed without authorization, could result in substantial harm, embarrassment, inconvenience, or unfairness to an individual or employee. PII includes name, address, email, telephone number, date of birth, SSN, and any other information that can be used alone or in conjunction with other information to identify an individual.

In academic research, one can take proper steps when designing research studies to avoid issues of PII entirely (e.g., simply not collecting it); when PII does need to be collected, there exists an institutional review board (IRB) to approve it. In non-academic settings, a division is typically responsible for reviewing and approving certain types of material (e.g., surveys, web pages). These divisions review material or processes that are considered at risk for PII breach and, like an IRB, can add considerably to project timelines. In applied settings, PII exists in virtually all work. Steps often need to be taken to ensure data are kept only on internal servers (with limited access), password protected, follow organization-specific policies for protecting employee PII, and comply with privacy laws and regulations (e.g., the European Union's General Data Protection Regulation [GDRP]; see Mintern & Rayner, 2018 for a review of GDRP compliance for employee data).

Proprietary Data

Theory and applied research both have value and typically complement each other; however, marrying the two for publication and dissemination has challenges and obstacles. Some non-academic settings consider information and data proprietary – unable to be shared in the public domain even when they are summarized in groups or anonymized. This approach to information may be frustrating for some practitioners, as it limits the ability to publish or present important applied research. There are various reasons for the protective stance on data, including protection from information that may create a negative public perception (e.g., high turnover rates, low-scoring attitudinal data on job satisfaction) and protecting competitive advantages from industry peers (e.g., hiring practices, training). Thus, the important and interesting research that practitioners conduct may not be able to be shared in the public sphere or with other researchers.

In our experience, this tends to be true more in the private sector and industry than the public sector (e.g., government). Even in organizations that do allow their members to speak, present, or publish, permission to do so may need to be given by the organization; disclaimers indicating that the views or opinions of the presenter or researcher do not represent those of the organization are common. If conducting research and publishing is important, it is often helpful to ask about and consider policies regarding conducting research and publishing before working in a specific applied setting.

In sum, there are several obstacles that practitioners may encounter pertaining to data and data analysis, many of which can complicate work and impede the ability to conduct and share research. While these limitations do exist, there are steps practitioners can take to work around these obstacles. First, with respect to how data are recorded and stored, we recommend involvement in any initiative that directly impacts your work. Attempts to have a voice in what is captured in any new system being implemented (e.g., database, software, applicant tracking system) should be made whenever possible. This may involve a significant amount of time and not be directly related to one's job description or duties. However, more ownership and involvement over the data collected can facilitate future data analysis and research efforts.

Additionally, if a system lacks data needed for research, there are alternative (albeit more time-consuming) ways it can be collected or mined. For example, referring back to the structured interview example, the original paper or electronic rating forms could be retrieved, and the data manually entered. Regarding publication policies in applied settings, it can be helpful to communicate and frame publishing as beneficial for the employer. For example, publishing can help employees build up their resumes; this can increase staff expertise and credibility. Other benefits that may persuade employers to allow publishing include providing more visibility to the organization and being able to advertise and emphasize career development when recruiting new staff.

Statistical and Analytical Approaches

Statistical analysis is an important tool for practitioners. The type of statistical analyses used by practitioners varies by project, but rudimentary analyses tend to be more commonly used than advanced analyses. For many research projects, frequencies and tables showing cross-tabulations are the most commonly reported statistics. In our experience, there are a number of statistical analyses that we have used regularly (e.g., correlation coefficients, frequencies, descriptive statistics, reliabilities), occasionally (e.g., chi-square, *t*-tests, analysis of variance, multiple regression), rarely (e.g., multivariate analysis of variance, non-parametric statistics, missing value analysis, meta-analysis), or never (e.g., cluster analysis, hazard analysis, time-series analysis). However, we do think it is important to have training in as many statistical techniques as possible and to be able to refamiliarize with them, as needed. To illustrate which statistical techniques are used by practitioners and for which activities, we obtained a list of common areas of specialization for applied social and behavioral scientists (Training Industry, 2020). Next, we linked each of the activities to major statistical analyses covered in our undergraduate and graduate coursework and from statistical textbooks. The results are presented in Table 37.1 and identify which statistical analyses are used for different areas of applied work.

Table 37.1a *Statistical techniques used in applied settings by area of specialization*

Area of specialization	Power analysis	Frequencies	Missing data analysis	*t*-test	Effect size	Chi-square	Bivariate/multiple correlation	Linear/logistic regression
Testing/assessment	X	X	X	X	X	X	X	X
Job analysis/job design/competency modeling	X	X	X	X	X	X	X	
Data analysis/research methods/statistics	X	X	X	X	X	X	X	X
Coaching/leadership/leadership development	X	X	X	X	X			
Performance management/succession planning/talent management	X	X	X	X	X		X	X
Motivation/job attitudes/engagement/surveys	X	X	X	X	X	X	X	X
Training/training and development	X	X	X	X	X		X	X
Organizational development	X	X						
Organizational performance/organizational effectiveness/change management/downsizing	X	X						
Strategy/strategic human resources/strategic human capital	X	X						
Recruitment/talent acquisition/sourcing	X	X	X	X	X	X	X	X
Inclusion/diversity	X	X	X	X	X	X	X	X
Careers/career planning/mentoring/socialization/ onboarding/ retirement	X	X						
Groups/teams	X	X	X	X	X		X	X
Human resources technology	X	X						
Legal issues/employment law/equal employment opportunity	X	X	X	X	X	X	X	X
General human resources	X	X						
Workforce planning	X	X	X	X	X	X	X	X
Cross-cultural/global issues	X	X	X	X	X	X	X	X
General organizational behavior	X	X	X	X	X	X	X	X
Work–family/work–life balance	X	X	X	X	X		X	X
Worker well-being/occupational health/safety/stress and strain/aging	X	X					X	

Table 37.1b (*cont.*)

Area of specialization	ANOVA or ANCOVA	MANOVA or MANCOVA	PCA/EFA/ CFA	Canonical correlation	Discriminant function analysis	Survival analysis	Time-series analysis	Structural equation modeling
Testing/assessment	X		X					X
Job analysis/job design/competency modeling	X							
Data analysis/research methods/statistics	X	X	X	X	X	X	X	X
Coaching/leadership/leadership development								
Performance management/succession planning/talent management								
Motivation/job attitudes/engagement/surveys	X	X	X		X			X
Training/training and development	X				X			
Organizational development								
Organizational performance/organizational effectiveness/ change management/downsizing	X							
Strategy/strategic human resources/strategic human capital								
Recruitment/talent acquisition/sourcing	X							
Inclusion/diversity								
Careers/career planning/mentoring/socialization/ onboarding/ retirement								
Groups/teams								
Human resources technology								
Legal issues/employment law/equal employment opportunity	X							
General human resources								
Workforce planning	X				X	X	X	
Cross-cultural/global issues			X					
General organizational behavior								
Work–family/work–life balance								
Worker well-being/occupational health/safety/stress and strain/ aging								

Table 37.1c

Area of specialization	Classical test theory	Item response theory	Generalizability theory	Profile analysis	Multilevel modeling	Multiway frequency analysis/log-linear analysis
Testing/assessment	X	X	X			
Job analysis/job design/competency modeling	X		X			
Data analysis/research methods/statistics	X	X	X	X	X	X
Coaching/leadership/leadership development						
Performance management/succession planning/talent management	X		X			
Motivation/job attitudes/engagement/surveys	X				X	
Training/training and development	X					
Organizational development						
Organizational performance/organizational effectiveness/change management/downsizing						
Strategy/strategic human resources/strategic human capital						
Recruitment/talent acquisition/sourcing						
Inclusion/diversity						
Careers/career planning/mentoring/socialization/ onboarding/retirement						
Groups/teams					X	
Human resources technology						
Legal issues/employment law/equal employment opportunity	X					
General human resources						
Workforce planning						
Cross-cultural/global issues						
General organizational behavior						
Work–family/work–life balance						
Worker well-being/occupational health/safety/stress and strain/aging						

Areas of specialization were taken from Training Industry (2020). ANOVA: analysis of variance; ANCOVA: analysis of covariance; MANOVA: multivariate ANOVA; MANCOVA: multivariate ANCOVA; PCA: principal components analysis; EFA: exploratory factor analysis; CFA: confirmatory factor analysis.

There are some occasions when more specialized and advanced statistics are needed, and sometimes practitioners need to learn a new area of statistics. For instance, the authors have been involved in conducting studies of employees across different locations and specialties. This required creating a stratified random sample and learning a weighting analysis, involving multiple stratifying variables, to estimate the overall results after accounting for oversampling of certain groups. Another study involved comparing two types of raters for an assessment and required learning how to calculate several measures of inter-rater agreement and consistency.

Other analyses the authors have used frequently include item analyses (e.g., computing *p*-values, item-total point-biserial and biserial correlations, distractor analyses) and reliability estimation (e.g., inter-rater reliability, Kuder–Richardson Formula 20, coefficient alpha, omega coefficients). Exploratory and confirmatory factor analyses and principal components analyses are sometimes conducted when there is a need to investigate the construct validity of scale scores. For example, the authors were involved in a job satisfaction survey research program in which organizational leaders and stakeholders noticed that the scores on different scales tended to increase or decrease in unison with each administration. This led the authors to conduct factor analyses on the data; they found evidence of a large general factor that influenced the results (Berger et al., 2015; Cucina & Byle, 2014; Cucina et al., 2014a).

In personnel selection research, regression analysis is often used to determine if different predictors add incremental validity over one another. Additional analyses used in applied work include computing means, standard deviations, Cohen's *d* effect size (Cohen, 1992), pass rates, and adverse impact ratios for different demographic groups. Power analysis also plays an important role in applied research, as a sufficient number of cases need to be obtained for there to be adequate power in a study (Schmidt et al., 1976). The authors have planned a number of studies by obtaining estimates of uncorrected validity coefficients from the literature and using G*POWER (Faul et al., 2007, 2009) to conduct a power analysis (see Chapter 6 in this volume).

Data analysis tools that are available to practitioners in non-academic settings may vary considerably. Depending on the industry and funding, researchers may only have access to one software platform, which may not match one's preference. Additionally, the cost of certain software may not outweigh the benefit of having a license to use it, or only a limited number of licenses may be issued. In our experience, the Excel and SPSS statistical analysis programs are utilized most frequently. An advantage of Excel is the ease of which data can be viewed, sorted, filtered, manipulated, and entered. In addition to being able to easily create graphs and charts, it has basic programming capabilities in terms of formulas and more advanced capabilities for developing macros.

Excel can also interface with numerous other programs, including Tableau, Power BI, and PowerPoint. In addition, Excel is useful for implementing psychometric and statistical formulas that are not in a statistical program, and it can be used to create syntax files for other statistical software (e.g., R, SAS, or SPSS). Additionally, many practitioners and stakeholders are familiar with using it, making sharing data sets and results with them much easier. With respect to software, we advise practitioners to learn how to conduct commonly used statistical analyses (e.g., *t*-test, correlations, regressions) in as many platforms as possible, and to develop basic fluency in Microsoft Excel, SPSS, SAS, and R.

Applied Setting Challenges

Certain aspects of working in applied settings are uniquely different than academia. Organizations are very interconnected, and a chain-of-command or hierarchy exists through which all work flows. The structure of organizations can often determine how the work is carried out, supervised, and prioritized. Practitioners also have several different customers and stakeholders. Work conducted often affects individuals both internal (e.g., employees, job incumbents) and external (e.g., job applicants) to the organization. Based on our experiences, we discuss several challenges practitioners encounter and make suggestions and recommendations for how to navigate these situations.

High-Stakes Environment

Because much applied research is used to make decisions about job applicants and employees, perhaps one of the greatest challenges that practitioners experience is the high-stakes environment. The high-stakes experience in applied settings tends to be at the group level. That is, if a mistake is made, it may affect both you and the entire organization. Academics experience a different kind of high-stakes environment that tends to be more on the individual level (e.g., publish or perish).

High-stakes decisions can be present in several applied situations. Personnel selection and promotional systems are likely the best examples of high-stakes work. An organization could find itself in violation of Title VII of the Civil Rights Act of 1964 if it is using a selection system that is not properly developed and validated (e.g., has substantially different pass rates for protected classes – race/ethnicity). Training programs can also fall under the high-stakes umbrella if individuals who do not successfully complete training are demoted or fired. If the failure rates for a training program differ by protected group status, the training program is treated akin to an employment test according to the Uniform Guidelines.

Similarly, performance management systems that lead to differences in pay, promotions, or dismissals may also come under scrutiny. Essentially, an employer needs to demonstrate the job-relevance of any employment decisions that have an adverse impact against protected groups. The process for demonstrating the job-relevance of employment decisions is described in the Uniform Guidelines; these are used in the enforcement of the Civil Rights Act of 1964 and the Equal Employment Opportunity Act of 1972 for employers in the United States. Other countries have similar laws and guidance (e.g., the Race Equality Directive 2000/43/EC and Equality Framework Directive 2000/78/EC in the European Union, guidance from the Equality and Human Rights Commission, 2014, in Great Britain).

Finally, emerging technologies (e.g., artificial intelligence and machine learning) are being increasingly applied in the workplace. Artificial intelligence and machine learning are being used for robotic process automation of certain tasks (e.g., setting up meetings, sending out reminders), hiring and selection (e.g., prescreening, interviewing, applications reviews), and predicting turnover potential of current employees. The wide-ranging application of artificial intelligence and machine learning has resulted in these technologies being used and applied in the workplace by those who have very little (or no) training in assessing outcomes for bias and disparate impact. It is critical that practitioners have an awareness of how artificial intelligence and machine learning are being utilized in the workplace, so they can properly develop and validate these tools when used in high-stakes situations. Tippins et al. (2021) describe several aspects of these technologies that practitioners should be vigilant of when applying them in the workplace.

Practitioners have an ethical and professional responsibility to be diligent in their work and to not discriminate (intentionally or unintentionally) on the basis of protected group status against applicants, trainees, and employees. When carrying out applied work, it is crucial to be mindful of the financial and legal implications that could result for the organization, especially when employment decisions are incorrect or not legally defensible. An organization may also experience negative publicity, and this could impact its bottom line or ability to conduct its mission. A key role for a practitioner is to be able to convey the short- and long-term benefits and risks of different high-stakes practices to an organization's leaders. Involving legal counsel can be helpful in this situation; however, sometimes an attorney specializing in employment law needs to be involved or assistance may need to be provided to find and share relevant court cases, laws, and regulations. Final decisions on programs and policies are sometimes not within the sole control of a practitioner, and undesirable outcomes can happen, especially in a litigious environment.

To protect the organization from damaging legal actions, one approach that many practitioners take is writing very detailed (and sometimes quite lengthy) technical reports about any system or process that is used to make decisions about job applicants or employees. For example, very thorough technical reports are written for job analysis and validation studies that are conducted for pre-employment tests; these include defensibility, adverse impact potential, and validity. In general, if the technical documentation is of exceptional quality, a plaintiff's team of lawyers and expert witnesses will often advise against pursuing a challenge in court, as the likelihood of winning is low. It is also important for practitioners to be aware of recent developments in the relevant literature, statistical tests, the assumptions underlying the tests, case law, any changes to relevant laws or regulations, and the potential arguments and criticisms that a plaintiff's expert witnesses might use when challenging a program.

Scope of Work

There are differences in the focus and scope of work for practitioners and academics. Academia offers a level of intellectual freedom that few other jobs have. Academics are very autonomous – they are able to conduct research they are passionate about, teach classes with minimal oversight, and somewhat set their own schedule. Individuals working in academia tend to focus more on research and activities that benefit the field as a whole. For example, an academic researcher might conduct a meta-analysis of the relationship between two constructs or design complex research studies to test hypotheses, and increment toward scientific theory development. Work in non-academic settings can be very different and offers several stark contrasts to academia. In non-academic settings, practitioners work with other employees towards a common vision or direction. Practitioners tend to focus their work on solving specific business and workplace problems and supporting the creation of a productive and effective workforce (e.g., through selection, training, performance management, teambuilding, and leadership development).

The individual versus shared vision contrast of work often manifests itself in the type of research and work conducted. In academia, there tends to be more freedom to choose research topics, whereas a practitioner often must focus on the research topics that are currently of greatest interest to the organization or that can be studied with available data. Thus, practitioners are somewhat more confined by an organization's mission, strategy, goals, available data, or participants than an academic researcher. While projects are typically matched up with skill sets, a practitioner cannot always control the direction of research or programs. Projects are sometimes selected based on a need instead of questions you choose to seek answers to. Furthermore, certain projects may be assigned that practitioners have low interest in or that are even outside the scope of the position. While these drawbacks may exist, the benefits of working in applied settings are that a plethora of data is available to access and use, and practitioners can see the immediate impact their research has on the workforce in real time. Those working in academia experience long publication timelines, and the impact of research may be less visible.

Time Pressures and Work Oversight

In academia, personal research and other projects tend to be largely within one's own control and sphere of influence, and the timeline can be largely self-determined. Furthermore, research teams tend to be small and manageable, and there are accessible samples to carry out research (e.g., constantly replenished introductory class subject pools). Occasionally, there are challenges that arise, such as a need for external funding (e.g., grants), special population samples, or a back-and-forth process with an IRB. Additionally, tenure-track academics may have aggressive timelines for accomplishing milestones and publishing work. These characteristics of research in academia lead to a high degree of predictable and controllable time frames and outcomes determinable mostly by self-management. In contrast, applied work has a much higher degree of interdependence, as individuals from many groups or areas of an organization are typically involved, and many outcomes are dependent on cooperation and timely responses from them.

Project timelines in applied settings are also constantly monitored for progress by management, and there are reporting requirements. A project plan will often be created for extensive projects, and updates will be given to leadership periodically, along with progress updates and justifications for why the project might be off-track. Oftentimes, pressure exists to deliver products or complete projects as quickly as possible, and stakeholders can greatly underestimate the time involved in conducting a study. Work projects are often part of yearly performance goals and plans for yourself and your superiors. Practitioners may have to adjust work timelines, regardless of the appropriate amount of time that may be necessary to complete it thoroughly. Furthermore, while working on multiple projects, other smaller projects or emergency work may come up unannounced – typically with short completion turnaround times. This can impact work in other areas, and it makes managing projects and research challenging. Other times, funding might only be available for a specific amount of time, or a contract might have hard deadlines, so there often is a sense of urgency in carrying out work.

There is also a high degree of oversight on project work in applied settings that can result in decision making that is not completely within your control. Practitioners should expect that work will be reviewed and approved by several superiors before it is accepted. At times, decisions will be made that you disagree with or do not follow theory or data – you must accept or proceed with it. For work that you are the final decision maker on, there are situations where you will have to decide between two options or alternatives that are not ideal. In short, work and ideas may not always conform to theory, and decisions may not always be data-driven.

Finally, there are also organizational decisions that can significantly affect aspects of work. There are times when resources needed to accomplish certain work will be reduced or limited. Resource reduction can come in several forms (e.g., staffing levels and budget limitations), and reductions in these areas may be limiting and inadequate to support and maintain the amount of work practitioners are asked to do. Organizational restructuring can also happen such that a practitioner's reporting chain is affected and even job duties shifted or changed; this can result in changes in scope of work or responsibility.

Recommendations

There are several actions practitioners can take to navigate these challenges and situations. Regarding time pressures, shifting priorities, and deadlines, we recommend that individuals in applied work embrace change by becoming as flexible and adaptable as possible. We also recommend that individuals learn to actively manage expectations of others. It has been our experience that explaining the steps and work required for projects, as they translate to staffing requirements or full-time equivalents, helps others to understand the workload required for projects. Furthermore, it may be beneficial to identify and communicate the operational, administrative, and maintenance work that is needed to carry out certain programs.

For example, if a structured interview is designed to help hire for an occupation, there is a tendency to focus predominately on what is required to create and implement the structured interview. What is more difficult to estimate and communicate is the level of ongoing work that will be required to maintain the structured interview program. In this instance, there is additional work required to ensure the program is run effectively, such as training staff how to use the tool and interview within legal guidelines, maintaining a list of certified interviewers who have completed that training, responding to challenges to hiring decisions, and updating interview forms over time.

Critical Skills and Duties in Applied Settings

Perhaps one of the largest differences between work in academic and non-academic settings is that there are skills needed in several areas beyond one's area of expertise, which are necessary to perform work in applied settings. In both settings, content expertise (e.g., psychology, methodology, statistics, psychometrics) is especially important, particularly at the individual contributor level (e.g., applied social scientist, professor at a university). As a practitioner, there are skills that you perform for work, such as literature reviews, research, data analysis, report writing, and interpreting and communicating the results of research. However, there are several more general skills, many of which are related to administration and management, that are often needed beyond one's academic training or area of expertise. Some of these skills are dependent on the industry and your role in the workplace. In this section, we describe some of the most common skills practitioners can develop to ensure successful work in applied settings.

Project Planning and Management

The degree of complexity in project planning and management is very different between academic and non-academic settings. In academia, research projects tend to be siloed, most resources needed to carry out research are accessible, and input is not needed from multiple groups of people to complete it. Research projects in non-academic settings tend to be more complex, as they require commitment and support from several groups across the organization, involve a cross section of employees, and often require resources (e.g., time from employees, funding) that one does not have direct control over. In this section, we describe specific project management skills needed to successfully carry out research in applied settings as well as how to navigate common obstacles.

First, skills in planning and forecasting timelines are critical to project management, as practitioners are often asked when products will be finished. There is generally a degree of unpredictability in planning and forecasting work in applied settings – many extraneous factors and influences can impact projects and timelines. Costs (e.g., salary for those involved, travel, equipment, materials, services, and who might provide funding), scheduling (e.g., whether the project might conflict with participants' busy seasons), deconfliction with other work, union involvement, management approvals, and issues facing the organization (e.g., downsizing) are some common examples of obstacles that can impact the ability to accurately determine how long work will take to complete. Awareness of the structure of the workplace (e.g., approval chains, how to procure equipment, who to go to for what) and how it operates is vital to planning and forecasting work and research; this often takes a considerable amount of time to learn. Senior employees with project management experience can help practitioners develop project planning skills on the job, and we recommend consulting with them to help create realistic project plans, incorporate organizational specific requirements, and anticipate risks.

Maintaining project planning documents to guide work and research is also common in most applied settings. The exact nature and format of these documents can vary by setting, but they typically include a narrative description of the project and a timeline. Timelines for many applied research projects can be created in Microsoft Word and Excel. More advanced project planning software is available (e.g., Microsoft Project); however, in our experience, that software is often more powerful than what is needed for typical applied research. Another common format for project planning is a Gantt chart. A Gantt chart contains bar charts that graphically illustrate the project schedule and each project step. Gantt charts can help break the project into smaller steps with anticipated start and completion dates and show dependencies and relationships for each step to stakeholders. It is often helpful for practitioners to familiarize themselves with what documents are used to create project plans where they are working. Additionally, training on project planning is available through many outlets for those new to applied settings (e.g., the Project Management Institute).

During the project itself, there will inevitably be details that were not anticipated ahead of time and other obstacles that arise during execution. Monitoring and communication skills are often critical at this phase, and the research staff should be prepared to provide regular updates to leadership on the progress of the project. Sometimes, conflicts, delays, and participation issues need to be escalated to the attention of leadership, and it can be helpful to have a designated point of contact on the project team or a champion in senior management to address these issues.

Program Management and Administration

Program management and administration – the process of planning and organizing products, documents, and processes for an initiative – are other areas practitioners dedicate time to when working in applied settings. When a practitioner creates a product (e.g., survey, test, training), there is operational and maintenance work that accompanies it. For instance, once a training program is developed and implemented, a practitioner also needs to maintain the training program for the duration of its use. In this situation, examples of ongoing work that might need to be done includes activities, such as tracking and maintaining a list of all trainees, editing the training over time, creating new training exercises, serving as a representative or point of contact for the training program, briefing leaders who seek information on the training, responding to additional requests for training, and so forth.

In many ways, creating a product for an organization to use is not as difficult and time-consuming as implementing and maintaining it. Depending on the complexity of the operations and maintenance work, a practitioner may have to assume this work. If the maintenance work is highly complex, practitioners must do it because it cannot be delegated to a non-technical employee; maintenance work that is more administrative in nature (e.g., maintaining a list of certified trainees) can be delegated to a non-technical employee or intern. However, with the increase of technology and automation, the number of administrative support staff has decreased, and many companies view administrative assistants as an unnecessary cost. As a result, "1.6 million secretarial and administrative-assistant jobs have vanished since 2000" (Feintzeig, 2020); this represents a 40% decrease in these positions. Because of this, non-technical or administrative employees may not be available to delegate work to, and practitioners may have to conduct administrative work themselves. Additionally, depending on your industry and role, you may find yourself involved in or leading work outside your area of expertise; for example, contracting can include supervising the work of contractor performance, keeping documentation of the contract, and evaluating final deliverables.

Organizational Communication

There is a large difference between academic and business writing and communication in both style and tone. Academics and practitioners largely have different audiences. The primary audiences in academic writing and communication are informed audiences, such as students via course instruction and peers through research and publications. Thus, writing and communication tends to take a formal approach. Practitioners communicate with varied audiences and communicate very different messages and for different purposes. Sometimes, practitioners communicate with a technical audience (e.g., peers in the field or organization), such as when communicating methodology, results, or outcomes of a study. Most often, however, practitioners communicate with a non-technical audience or those in other professions. Audiences in applied settings include legal counsel, executives, information technologists, and employees specific to the industry you are working in (e.g., finance, healthcare, law enforcement). Because of the various situations, it is difficult to address each and every scenario that occurs in the workplace. However, there are a few overarching guidelines that we view as helpful to consider when communicating verbally or in writing in the workplace:

- Messaging often involves communicating a process or policy. Practitioners should use very process-oriented language or refer to standard operating procedures when possible.
- Communicate ideas clearly and concisely; sometime bullets or "one-pagers" are all a leader has time to read.
- Oftentimes, employees want to know the "why" behind actions, so rationales should be included, as necessary.
- Practitioners are asked to solve problems, so recommending a course of action is preferred.

- On important messages that are sent across the organization or to executives, peers and superiors should review the message before sending, to ensure the correct tone, interpretation, content, and agreement with messaging.
- Email messages and documents created at work (e.g., technical reports) are "records," stored indefinitely, and can be used in legal proceedings; careful wording is often necessary.

Deciding What is Best for You

We have addressed various topics throughout this chapter and described the most common challenges and differences of applied work compared to academic work, but how should an individual approach the decision-making process regarding whether applied work is a fit? What questions should you ask yourself (or others) when deciding between academic and practitioner career paths? We conclude the chapter by presenting several steps an individual can take when choosing a career path. At a high level, we recommend that an individual take an analytical approach. As someone pursuing or possessing an advanced degree, the analytical skills necessary to complete that process can be helpful to thoroughly evaluate choices such as this. We advise others to use those analytical skills and list out the advantages and disadvantages, given some of the considerations we have suggested in this chapter. We also recognize that decision making for pursuing academic or applied work is a very personal choice, and there are factors beyond work that influence one's decision.

The most obvious consideration when deciding whether applied work is a proper fit is the environment. Academic and non-academic settings operate very differently, and an individual seeking applied work should be comfortable with this difference. A logical first step to take in understanding work outside academia is to talk to people from industries you are interested in. Individuals working inside organizations often have the best perspective and can provide the most realistic job preview. This is an especially important step because there is a wide variety of industries to work in (e.g., law enforcement, insurance, consulting), and there are several factors to consider beyond the technical aspects of work (e.g., the subject matter). Related to this, there are different sets of job responsibilities and duties for jobs across work settings – each with their own advantages and disadvantages. It would be impossible to describe or summarize, in this chapter, the differences among the various types of applied workplaces. Thus, the most informative way to determine whether applied work is the right fit is to try to understand a specific situation before entering it.

A second consideration an individual should make is career path and trajectory. Because of the nature of the positions in academia and applied settings, career progression often has a different path and trajectory, and promotion and tenure can vary. For a practitioner, completion of operational project work is often critical for promotion, with research and years of service playing less of a role (although there are exceptions). In contrast, in academia, applicants for tenure-track positions are often expected to give a job talk and have multiple interviews with faculty, students, and a search committee. Tenure and promotion can hinge heavily on publication record, followed by teaching, and then service. The concept of tenure is largely non-existent in applied settings, although in some organizations (e.g., the civil service), employees may have to undergo a probationary period after which time it is much more difficult for them to be removed due to poor performance. Practitioners working as internal consultants in the private sector, or for consulting firms that rely on incoming revenue, are not immune to layoffs that can occur with economic downturns.

Finally, while there are variances among different workplace settings and career paths, there are steps one can take as an individual to assess the self. We recommend assessing personality and preferences in an honest and candid manner, either formally or informally, particularly with respect to the situations we have described in this chapter. Formally, there are personality instruments that assist with career and vocational choice, such as the Strong Interest Inventory (Harmon et. al., 1994). Although we do not necessarily feel that these instruments are determinative with respect to career choice, they may help individuals begin the process of assessing how well their personality, interests, preferences, and habits match with applied work and settings. Informally, we have summarized important work characteristics, many of which we mentioned in this chapter, that tend to distinguish applied work from academic work. These characteristics should also be considered in conjunction with one's personality, interests, preferences, and habits, and are listed and summarized in Table 37.2.

Lastly, it is important to note that there are various combinations and balances of work that can be achieved. There are situations in which one works primarily in academia while conducting occasional applied work and others in which one

Table 37.2 *Academic and applied/industry characteristics*

Characteristic	Academic	Applied/industry
Research		
Research dissemination	**Intellectual freedom:** You are able to publish all research	**Proprietary:** Research often cannot be shared or published due to industry competition; some exceptions exist (e.g., government)
Research–interest match	**High:** You are able to work exclusively on research matched to interest	**Variable:** You tend to do some research in areas of interest; other work assigned may be of low interest
Research outcomes	These tend to be focused on theory; the focus is on publications and citations; there may be long times between research and publication; you are often unable to see impact outside academia	These tend to be focused on workplace outcomes; organizational decisions are made based on research; you can see results and impact of research in real time

Table 37.2 (*cont.*)

Characteristic	Academic	Applied/industry
Work structure:		
Work style	**Individual/solitary:** Research and teaching at a university tends to be more individual and solitary	**Teamwork-oriented:** Most work is done in teams; you frequently conduct interdisciplinary work
Workplace role stability	**High:** Changes in departmental stability rarely occur; there are clear expectations to perform the same role over time	**Variable:** Organizational restructuring occurs; you may have shifts in roles and responsibilities over time
Reporting structure	**Unstructured:** Typically you have no direct reports; you teach and work with no or little supervision	**Highly structured:** You may have many direct reports; several layers of management exist; reporting of work is done on a frequent basis (e.g., weekly)
Supervision responsibilities	You supervise graduate assistant work and research	You lead interdisciplinary project teams; you may manage or supervise groups of employees
Other:		
Work–life balance	**Achievable:** You have summers off or with reduced teaching responsibilities; sabbaticals; there is flexibility of schedule in some instances	**Achievable:** Most work typically conducted during business hours (8am–5pm); work from home options, telecommuting, and alternative work schedules are sometimes available
Pay/ compensation	This tends to be lower than applied work/industry; there is high job stability	This tends to be higher than academia; there is lower job stability depending on the industry
Promotion and development	Focus is on achieving tenure; some promotional opportunities are available in college administration	More upward mobility exists, especially if you are willing to work outside field of interest

conducts primarily applied work while adjunct teaching – an individual can have a foot in both worlds. A common question people have is whether a person can switch mid-career between applied and academic work. While doing so is possible, there may be implications for career trajectory. For academic jobs, the focus is on eventually achieving tenure, a process that usually takes a significant amount of time and academic achievements (e.g., publication record) to build up to. We feel the process of going from academic work to applied work is easier than going from applied work to academic work. This is because, in the private and public sectors, there are generally a lot of opportunities for promotion, especially if an individual is willing or interested in transitioning to a role outside of one's area of expertise.

References

Aguinis, H. & Pierce, C. A. (2008). Enhancing the relevance of organizational behavior by embracing performance management research. *Journal of Organizational Behavior*, *29*, 139–145. https://doi.org/10.1002/job.493

Anderson, N. (2007). The practitioner–researcher divide revisited: Strategic-level bridges and the roles of IWO psychologists. *Journal of Occupational and Organizational Psychology*, *80*, 175–183. https://doi.org/10.1348/096317907X187237

Berger, J. L., Cucina, J. M., Walmsley, P. T., & Martin, N. R. (2015). General factor in employee surveys: A large-sample investigation. Poster presented at the 30th meeting of the Society for Industrial and Organizational Psychology, Philadelphia, PA, 23–25 April.

Cascio, W. F. (2008). To prosper, organizational psychology should . . . bridge application and scholarship. *Journal of Organizational Behavior*, *29*(4), 455–468. https://doi.org/10.1002/job.528

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cook, T. D. & Campbell, D. T. (1979). *Quasi Experimentation: Design and Analytical Issues for Field Settings*. Rand McNally.

Cucina, J. M. & Byle, K. A. (2014). *Technical Note: The Federal Employee Viewpoint Survey (FEVS) Measures a Large General Factor (and a Few Smaller Ones*). US Customs and Border Protection.

Cucina, J. M., Credé, M., Curtin, P. J., Walmsley, P. T., & Martin, N. R. (2014a). Large sample evidence of a general factor in employee surveys. Poster presented at the 29th meeting of the Society for Industrial and Organizational Psychology, Honolulu, HI, May 15–17.

Cucina, J. M., Hayes, T. L., Walmsley, P. T., & Martin, N. R. (2014b). It is time to get medieval on the overproduction of pseudotheory: How Bacon (1267) and Alhazen (1021) can save I/O psychology. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *7*(3), 356–364. https://doi.org/10.1111/iops.12163

Equality and Human Rights Commission (2014). What equality law means for you as an employer: When you recruit someone to work for you. Equality Act 2010 Guidance for Employers. (Vol. 1). Available at: www.equalityhumanrights.com/sites/default/files/what_equality_law_means_for_you_as_an_employer_-_recruitment.pdf.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191. https://doi.org/10.3758/BF03193146

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Feintzeig, R. (2020). The vanishing executive assistant. *Wall Street Journal*, January 18. Available at: www.wsj.com/articles/the-vanishing-executive-assistant-11579323605.

Gelade, G. A. (2006). But what does it mean in practice? The *Journal of Occupational and Organizational Psychology* from a practitioner perspective. *Journal of Occupational and Organizational Psychology*, *79*, 153–160. https://doi.org/10.1348/096317905X85638

Harmon, L. W., Hansen, J. C., Borgen, F. H., & Hammer, A. L. (1994). *Strong Interest Inventory: Applications and Technical Guide*. Consulting Psychologists Press, Inc.

Hodgkinson, G. P. (2006). The role of JOOP (and other scientific journals) in bridging the practitioner–researcher divide in industrial, work, and organizational (IWO) psychology. *Journal of Occupational and Organizational Psychology*, *79*, 173–178. https://doi.org/10.1348/096317906X104013

Johnson, J. G. (1976). Albermarle Paper Company v. Moody: The aftermath of Griggs and the death of employee testing. *Hastings Law Journal*, *27*(6), 1239–1262.

Mintern, T. & Rayner, S. (2018). 8 Aspects of GDPR compliance: A brief guide for HR functions. Available at: www.lexology.com/library/detail.aspx?g=fac839cf-e292-4452-b1d6-f87c81e81424.

Rynes, S. L. (2007). Let's create a tipping point: What academics and practitioners can do, alone and together. *Academy of Management Journal*, *50*, 1046–1054. https://doi.org/10.5465/AMJ.2007.27156169

Rynes, S. L. (2012). The research–practice gap in I/O psychology and related fields: Challenges and potential solutions. In S. W. J. Kozlowski (ed.), *Oxford Library of Psychology. The Oxford Handbook of Organizational Psychology, Volume 1* (pp. 409–452). Oxford University Press.

Rynes, S. L., Bartunek, J. M., & Daft, R. L. (2001). Across the great divide: Knowledge creation and transfer between practitioners and academics. *Academy of Management Journal*, *44*, 340–355. https://doi.org/10.2307/3069460

Schmidt, F. L. & Hunter, J. E. (2003). History, development, evolution, and impact of validity generalization and meta-analysis methods, 1975–2002. In K. R. Murphy (ed.), *Validity Generalization: A Critical Review* (pp. 31–66). Erlbaum,.

Schmidt, F. L., Hunter, J. E., & Ury, V. W. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology*, *61*(4), 473–485. https://doi.org/10.1037/0021-9010.61.4.473

Tippins, N. T., Oswald, F. L., & McPhail, S. M. (2021). Scientific, legal, and ethical concerns about AI-based personnel selection tools: A call to action. *Personnel Assessment and Decisions*, *7*(2), 1–22. https://doi.org/10.25035/pad.2021.02.001

Training Industry (2020). *Income & Employment Report*. Society for Industrial and Organizational Psychology.

US Department of Labor (2000). *Testing and Assessment: An Employer's Guide to Good Practices*. Employment and Training Administration, US Department of Labor. Available at: www.onetcenter.org/dl_files/empTestAsse.pdf.

Index