

Principles and Methods of Social Research

Fourth
Edition

William D. Crano
Marilynn B. Brewer
Andrew Lac

PRINCIPLES AND METHODS OF SOCIAL RESEARCH

Through a multi-methodology approach, *Principles and Methods of Social Research, Fourth Edition* covers the latest research techniques and designs and guides readers toward the design and conduct of social research from the ground up. Applauded for its comprehensive coverage, the breadth and depth of content of this new edition is unparalleled.

Explained with updated applied examples useful to the social, behavioral, educational, and organizational sciences, the methods described are relevant to contemporary researchers. The underlying logic and mechanics of experimental, quasi-experimental, and nonexperimental research strategies are discussed in detail. Introductory chapters covering topics such as validity and reliability furnish readers with a firm understanding of foundational concepts. The book has chapters dedicated to sampling, interviewing, questionnaire design, stimulus scaling, observational methods, content analysis, implicit measures, dyadic and group methods, and meta-analysis to cover these essential methodologies. Notable features include an emphasis on understanding the principles that govern the use of a method to facilitate the researcher's choice of the best technique for a given situation; use of the laboratory experiment as a touchstone to describe and evaluate field experiments, correlational designs, quasi-experiments, evaluation studies, and survey designs; and coverage of the ethics of social research including the power a researcher wields and tips on how to use it responsibly.

The new edition features:

- Increased attention to the distinction between conceptual replication and exact replication and how each contributes to cumulative science.
- Updated research examples that clarify the operation of various research design operations.
- More learning tools including more explanation of the basic concepts, more research examples, and more tables and figures, such as additional illustrations to include Internet content such as social media.
- Extensive revisions and expansions of all chapters.
- A fuller discussion of the dangers of unethical treatment to research participants.

Principles and Methods of Social Research, Fourth Edition is intended for graduate or advanced undergraduate courses in research methods in psychology, communication, sociology, education, public health, and marketing, and further appeals to researchers in various fields of social research, such as social psychology and communication.

William D. Crano is the Oskamp Distinguished Chair of Psychology and Director of the Health Psychology and Prevention Science Institute at Claremont Graduate University.

Marilynn B. Brewer is Professor Emeritus, Ohio State University.

Andrew Lac is an Associate Professor of Psychology at the University of Colorado, Colorado Springs.

PRINCIPLES AND METHODS OF SOCIAL RESEARCH

Fourth Edition

*William D. Crano, Marilyn B. Brewer,
and Andrew Lac*

Fourth edition published 2024
by Routledge
605 Third Avenue, New York, NY 10158

and by Routledge
4 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

© 2023 William D. Crano, Marilynn B. Brewer, Andrew Lac

The right of William D. Crano, Marilynn B. Brewer and Andrew Lac to be identified as authors of this work has been asserted in accordance with sections 77 and 78 of the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this book may be reprinted or reproduced or utilised in any form or by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying and recording, or in any information storage or retrieval system, without permission in writing from the publishers.

Trademark notice: Product or corporate names may be trademarks or registered trademarks and are used only for identification and explanation without intent to infringe.

First edition published by Allyn & Bacon 1986
Second edition published by Lawrence Erlbaum Associates, Inc 2002
Third edition published by Routledge 2015

Library of Congress Cataloging-in-Publication Data

Names: Crano, William D., 1942– author. | Brewer, Marilynn B., 1942– author. | Lac, Andrew, author.

Title: Principles and methods of social research / William D. Crano, Marilynn B. Brewer, Andrew Lac.

Description: Fourth edition. | New York, NY: Routledge, 2023. | Includes bibliographical references and index. | Identifiers: LCCN 2022060828 (print) | LCCN 2022060829

(ebook) | ISBN 9781032222417 (hardback) | ISBN 9781032222400 (paperback) | ISBN 9781003271734 (ebook)

Subjects: LCSH: Social sciences—Research. | Social sciences—Methodology.

Classification: LCC H62 .C692 2023 (print) | LCC H62 (ebook) | DDC 300.72—dc23/eng/20230105

LC record available at <https://lcn.loc.gov/2022060828>

LC ebook record available at <https://lcn.loc.gov/2022060829>

ISBN: 9781032222417 (hbk)

ISBN: 9781032222400 (pbk)

ISBN: 9781003271734 (ebk)

DOI: 10.4324/9781003271734

Typeset in Bembo
by codeMantra

CONTENTS

<i>Acknowledgments and Dedication</i>	<i>x</i>
<i>Preface</i>	<i>xi</i>
PART I	
Introduction to Social Research Methods	1
1 Basic Concepts	3
<i>Science and Daily Life</i>	<i>4</i>
<i>From Theory, Concept, or Idea to Operation</i>	<i>4</i>
<i>Role of Theory in Scientific Inquiry</i>	<i>13</i>
<i>Conclusion and Overview</i>	<i>14</i>
<i>References</i>	<i>16</i>
2 Internal and External Validity	18
<i>Causation</i>	<i>19</i>
<i>Distinguishing Internal and External Validity</i>	<i>22</i>
<i>Basic Issues of Internal Validity</i>	<i>27</i>
<i>Basic Issues of External Validity</i>	<i>35</i>
<i>Conclusion</i>	<i>37</i>
<i>References</i>	<i>38</i>
3 Measurement Reliability	40
<i>Classical Measurement Theory</i>	<i>40</i>
<i>Contemporary Measurement Theory</i>	<i>48</i>
<i>Conclusion</i>	<i>57</i>
<i>References</i>	<i>58</i>

4 Measurement Validity	60
<i>Types of Measurement Validity</i>	61
<i>The Multitrait-Multimethod Matrix: Combining Convergent and Discriminant Validity</i>	66
<i>Threats to Measurement Validity</i>	70
<i>Conclusion</i>	73
<i>References</i>	73

PART II

Research Design Strategies: Experiments, Quasi-Experiments, and Nonexperiments 77

5 Designing Experiments: Variations on the Basics	79
<i>Basic Variations in Experimental Design</i>	79
<i>Expanding the Number of Experimental Treatments</i>	82
<i>Block Designs: Incorporating a Nonexperimental Factor</i>	90
<i>Repeated-Measures Designs and Counterbalancing</i>	93
<i>Conclusion</i>	95
<i>References</i>	96
6 Constructing Laboratory Experiments	97
<i>Steps for Constructing an Experiment</i>	97
<i>Types of Experimental Manipulations</i>	100
<i>Manipulation and Attention Checks</i>	106
<i>Assignment of Participants to Conditions: Randomization Procedures</i>	107
<i>Realism and Engagement in an Experiment</i>	109
<i>Role-Playing Simulations and Analogue Experiments</i>	110
<i>Conclusion</i>	116
<i>References</i>	117
7 External Validity of Laboratory Experiments	120
<i>Generalizability across Participants</i>	121
<i>Experimenter Expectancy and Bias</i>	125
<i>Three Faces of External Validity</i>	130
<i>Conclusion</i>	133
<i>References</i>	133
8 Conducting Experiments Outside the Laboratory	136
<i>Research Settings and Issues of Validity</i>	137
<i>Constructing a Field Experiment</i>	138
<i>The Internet as a Site for Experimental Research</i>	144
<i>Conclusion</i>	150
<i>References</i>	151
9 Quasi-Experiments and Applied Research	155
<i>Quasi-Experimental Methods in Applied Contexts</i>	156

<i>Quasi-Experimental Designs</i>	163
<i>The Use of Archival Data in Longitudinal Research</i>	174
<i>Conclusion</i>	175
<i>References</i>	176
10 Nonexperimental Research: Correlational Design	179
<i>Bivariate Correlation and Regression</i>	181
<i>Multiple Regression</i>	190
<i>Using Regression to Test Mediation</i>	194
<i>Uses and Misuses of Correlational Analysis</i>	195
<i>Conclusions</i>	197
<i>References</i>	198
11 Advanced Multivariate Correlational Design	199
<i>Multilevel Models</i>	199
<i>Structural Equation Models</i>	202
<i>Modeling Longitudinal Data</i>	211
<i>Conclusion</i>	216
<i>References</i>	217
PART III	
Data Collecting Methods	219
12 Survey Studies: Design and Sampling	221
<i>Selection versus Assignment</i>	221
<i>Census and Survey Basics</i>	222
<i>Random Sampling</i>	226
<i>Non-random Sampling</i>	233
<i>Other Sampling Issues</i>	235
<i>Types of Survey Studies</i>	239
<i>Missing Data</i>	241
<i>Conclusion</i>	243
<i>References</i>	244
13 Systematic Observational Methods	248
<i>Three Aspects of Naturalism</i>	248
<i>Observer Involvement in the Naturalistic Setting</i>	251
<i>Coding Observations</i>	256
<i>Conclusion</i>	269
<i>References</i>	270
14 Content Analysis	274
<i>Content Analysis Basics</i>	274
<i>Conducting a Content Analysis</i>	275
<i>Summary of the General Paradigm</i>	284

viii Contents

Representative Examples 285
Conclusion 290
References 290

15 Interviewing 292

Modes of Administration 292
Developing the Interview 294
Conducting the Interview 301
Group Interviews and Focus Groups 308
Conclusion 310
References 310

16 Construction of Questionnaires and Rating Scales 314

Questionnaires 314
Constructing Rating Scales 318
Conclusion 329
References 330

17 Scaling Stimuli: Social Psychophysics 333

Scaling Stimuli 334
Stimulus Scaling Techniques 335
Multidimensional Scaling Models 340
Conclusion 345
References 345

18 Indirect and Implicit Measures of Cognition and Affect 348

Indirect Measures 349
Information Processing: Attention and Memory 350
Priming: Processing Without Awareness or Intent 356
Social Psychophysiology: Physiological Traces of Affect and Cognitive Processing 362
Conclusion 369
References 370

19 Methods for Assessing Dyads and Groups 374

Dyadic Designs 374
Designs to Study Group Structures 379
Designs to Study Multiple Groups 385
Measuring Group Process and Outcomes 387
Conclusion 390
References 391

PART IV**Concluding Perspectives 395****20 Synthesizing Research Results: Meta-Analysis 397***Replicability of Findings 397**Meta-Analysis 398**Stages in the Meta-Analysis Process 400**Interpreting the Meta-Analysis 406**Conclusion 414**References 415***21 Social Responsibility and Ethics in Social Research 417***Ethics of Research Practices 417**The Regulatory Context of Research Involving Human Participants 424**Ethics of Data Reporting 427**Ethical Issues Related to the Products of Scientific Research 429**Conclusion 430**References 431**Glossary 434**Author Index 447**Subject Index 463*

ACKNOWLEDGMENTS AND DEDICATION

This book is dedicated to the memory of our mentor, Donald T. Campbell, as have been all prior versions of this work. It is made in acknowledgment of his contributions to the social sciences, and of the debt we owe to him personally in fostering the development of our ideas about the right way to think about research. In his classes on research methods, Campbell never tried to teach what to do, or even how to do it, but rather how to think about questions and problems and how to approach their solution in a rigorous and scientifically sound manner. The continued evolution of research methods, designs, and statistics over the years attests to the astuteness of his teaching strategy. By focusing on the fundamentals that lie at the base of all research methods, students are made capable of absorbing new approaches as variations of what is already known. This allows continual updating of techniques and approaches without major dislocations. Following new developments is more a question of generalizing what is known to new contexts, rather than starting from scratch with each advance.

In addition to Don Campbell, many others have contributed to the development of this book. Our colleagues' constructive commentary on previous editions of the text have been considered, and sometimes even adopted. We appreciate these interchanges, as they undoubtedly allowed us to produce a work that is better than it would have been. We think this book has contributed to the scientific development of our students, and we know our students have contributed along the way to its continual refinement. To them, we owe our thanks. Finally, we owe a special thanks to Suellen, who has been with us from the very start of this journey, and who has contributed to our work in too many ways even to mention – we couldn't have done it without you.

PREFACE

When this book's "great-grandfather" first saw the light of day its authors were fresh out of graduate school, former students of Donald T. Campbell, one of social science's preeminent research methodologists. Then and now we recognized the need for a clear understanding of the proper application of the scientific method to the study of human social behavior. As before, many people today contend that science has nothing to say about the beliefs, values, and behavioral characteristics that define us as human. From this ultimately pessimistic point of view, love, hate, identity, alienation, prejudice, reason, discrimination, altruism, attitudes, intergroup relations, and a host of other distinctly human characteristics and properties are defined as beyond the reach of science. The argument goes, "Such things are better left to poets or priests – they are not appropriate subjects of scientific study." We disagree.

We disagree because we are convinced that in the short space of this book's life, good social research has added immensely to our store of knowledge about all of these difficult-to-research, but ultimately understandable issues. Complex aspects of human social life are subject to good scientific analysis, which when applied properly produces knowledge that leads to a better understanding of the human condition. This book is designed to help lay the necessary foundation for this quest. It is firmly grounded in the scientific method and provides the background knowledge necessary to design and implement good research. It is crucial, however, that the rigorous application of the scientific method is accompanied by creativity and dedication if real progress is to be made in social research. Although we believe this book can supply the necessary grounding, you will need to bring your own creativity and motivation.

In our earliest rendition of this book, we focused strongly on the experimental method as a touchstone against which other research methodologies were to be judged. We still view the experiment as the preeminent means of developing strong causal inferences, but even in the original volume we recognized that the environment of the laboratory was quite restrictive, and so we expanded the field of play to include topics like interviewing and content analysis, topics that at the time were deemed "esoteric" inclusions in a book on research design. In the book's later versions, we expanded our consideration of the more traditional experimental models, and extended our discussion to include new topics, including analogue experiments, meta-analysis, regression-based causal modeling, survey sampling, evaluation research, dyadic analysis, and experimenter expectancy effects. Emerging methods in social cognition also made the chapter

list, along with an augmented emphasis on external validity and generalizability, the use of field research accompanied by relatively novel techniques like propensity analysis and other methods of intervention and evaluation that fit with our growing realization that good science owes a debt to the society that nurtured it, a concern Campbell advanced insistently in his later years.

The basic structure of our book has been designed to build logically from fundamental research strategies (experimental and quasi-experimental methods, correlational designs, etc.) to the tactics of data collection and the logistics of proper research procedures. We have presented this material as clearly as we could and believe a conscientious student should not have problems understanding the concepts and advice included across the various chapters. To facilitate comprehension, important terms throughout the chapters are printed in boldface type, indicating a word or concept that has been included and defined in the glossary at the end of the book.

A unique feature of this book is its comprehensiveness. We do not know of a book of this type that considers such a broad range of experimental, nonexperimental, correlational, quasi-experimental, and evaluation research designs, and that has combined these with data collection methods ranging from survey sampling to interviewing to content analysis, scale development, and methods for assessing dyads and groups. There is much to do in developing a working understanding of the intricacies of good methodological design, and we believe that this presentation will facilitate its readers in doing so.

Content Overview

The volume is broken into four major multi-chapter sections. Part I is focused on the basics, the fundamental introductory concepts that must be well in hand if the rest of the material is to be properly learned and profitably employed. Crucial issues of reliability and validity occupy the initial section, as it should occupy the initial considerations of any research method. We emphasize the basics in the first chapter, the nature of the scientific method, operationalization, the place of theory in a developing science, and ideas of causation, moderation, and mediation. The second chapter leans heavily on Campbell and Stanley's classic exposition of threats to internal and external validity in experimentation. Chapter 3 is major reworking of measurement reliability theory. It includes discussion of classic and contemporary test theory, latent factors, exploratory and confirmatory factor analysis, and item response theory. In Chapter 4 we draw heavily on our earlier discussion of measurement validity. Here, we discuss the many faces of validity, the multi-trait multimethod matrix, and the various threats to validity that must be considered when developing trustworthy measures.

Part II is concerned with important features of experimental design. Experimentation is a fundamental feature of the book, as it provides a disciplined way of thinking about research that we believe is productive and generative. This mindset infuses our discussion of research methods that do not involve the strict control and constriction of the experiment. Chapter 5 is focused on true experimental designs, factorialization, and interpreting interaction effects. Chapter 6 is concerned with the working knowledge needed to design and conduct experiments. It deals with various forms of manipulations and manipulation checks, analogue and simulation studies, and mundane and experimental realism. Chapter 7 develops this latter theme; it highlights generalizability, or external validity, with a special emphasis on the nature of the individuals who serve as the participants in our studies, the possibly distorting effects of experimenter biases and expectancies on research results, and ways to avoid them.

In the later sections of Part II, we expand the canvas to consider research outside the friendly confines of the laboratory. Chapter 8 deals with the special issues involved in field experimentation, including research that makes use of the Internet as a source of both information

and research participants. Chapter 9 provides a discussion of quasi-experimental designs and evaluation research. This chapter is a unique entry not often included in many introductions to research design in the social sciences. However, these methods add value to the social scientist's methodological arsenal and should be well understood as they greatly expand the range of phenomena that can be investigated in a rigorous manner. Moving on to nonexperimental research methods, Chapters 10 and 11 provide extended coverage of correlational design and analyses, including considerations of multiple regression, path analysis, and latent structural equation approaches, which are used to specify and evaluate the adequacy of complex multivariate models involving nonexperimental data.

We noted in an earlier volume that there was never a single, *right* way to support a position. That being the case, it is important to have a command of a variety of methods beyond the experiment that allow different features of a phenomenon to be investigated in different ways. To hold out for the experiment over all other design alternatives suggests a shallow understanding of the research process, because some questions are *not* ideally studied with experimental methods, just as some almost demand them. To manipulate a research question so that it "fits" an experimental approach is backward. It is far better to fit the appropriate method to the question at issue, and often that method is not an experiment.

Part III of the book is concerned with data collection methods used to supply the raw material for the design possibilities discussed in Parts I and II. Here we learn about the many techniques in which grist for the social scientist's mill can be collected as reliably and validly as possible. Chapter 12 concerns the survey in all its various forms, with a focus on sampling methodologies. In addition, we consider various forms of missing data and their implications for interpretation of results. Chapter 13 provides an overview of systematic observational methods and ways to develop systems to code the often-unruly data of field observations. Chapter 14 deals with content analysis, and factors that must be considered when contemplating either a deductive or inductive form of communication analysis. Chapter 15 is focused on the various forms of interviewing, ranging from unsystematic exploratory interviews to highly structured, scheduled interviews, and the interviewing tactics designed to capture the most useful, reliable, and valid data from this method.

In Chapter 16, we focus on the proper design of items and responses for questionnaires and scales, using Thurstone's, Guttman's, Likert's, or Osgood's measurement approaches. Chapter 17 discusses psychophysical methods that have been in use in psychology since the 1800s. Psychophysical scales were among the earliest contributions of the pioneers of psychology, and include measures derived from pair comparison and rank order scaling. These techniques are valuable in and of themselves and serve as a lead-in to the more modern techniques of multidimensional scaling analysis. Chapter 18 is an exposition of techniques used in social cognition research to measure implicit thoughts and feelings. It describes the use of implicit approaches ranging from the Thematic Apperception Test to the Implicit Associations Test. The chapter also considers more physiologically oriented approaches, including event-related potential (brain waves), neuroimaging (positron emission tomography, or PET scans), and functional magnetic resonance imaging (fMRI).

Data describing dyadic and group interactions are the special focus of Chapter 19. This chapter outlines some of the ways in which this inherently social form of data can be understood, and the designs that can be used to capture this understanding. Round robin designs, sociometric assessments, and social network analysis are considered in this final chapter of Part III.

Part IV of the text consists of only two chapters, but both are crucial to the research enterprise. The first of these, Chapter 20, deals not with a specific research design method or technique, but rather with meta-analysis, a means of quantitatively synthesizing data across a host

of studies that have all focused on the same relationship. As a way of summarizing the collected knowledge of a phenomenon or variable relation, meta-analysis has a clear advantage over simple vote counting methods and narrative reviews in coming to conclusions about conceptual relationships. Fixed and random-effects models of meta-analysis are discussed, as are the potential effects of publication bias, the file drawer problem and how to assess it using approaches such as the funnel plot, and the various search strategies that are useful in any comprehensive literature search.

The final chapter of the book is devoted to a comprehensive consideration of the ethics of research. We considered a possible alternative approach, which involved spreading various ethical caveats thinly throughout the chapters but judged this to be a less useful strategy than a concentrated discussion of the ethics of social research in a dedicated chapter. This chapter provides a comprehensive treatment of ethical issues in research with human participants and considers the social responsibility that social researchers have to their participants, to science, and to society. Research ethics apply to both the conduct of research and the reporting and dissemination of research findings. With respect to data reporting, the chapter covers recent issues in the replicability of social research findings and the move toward open science and transparency in reporting methods and results of empirical studies.

Intended Audience

We believe this is one of the few books that is comprehensive enough to provide the necessary depth to prepare graduate and upper level undergraduate students to design, execute, critically evaluate, interpret, and describe various methodological approaches used in social research. This book is ideal for graduate or advanced undergraduate courses in research methods (design) in psychology, particularly social psychology, as well as in communication, sociology, education, social work, and public health. A previous course in research methods at the undergraduate level is useful in taking full advantage of this work, but we have found that with sufficient motivation, conscientious students in past years who did not have this prerequisite have been able to fare well.

New to This Edition

This brings us to the current version of this textbook. From the start, we set out to produce a book that focused almost exclusively on design issues, without digging into the minutia of statistical computations. We have maintained this focus, but the development of increasingly sophisticated methods of statistical analysis has opened the door to increasingly sophisticated research design strategies, and we have taken advantage of both in this latest edition. This edition contains a new chapter on multivariate analysis and structural equation modeling written in a way to introduce readers to the basic concepts and logic of these techniques without the computational details (that are best left to statistics books). Also new to this edition is extended attention to the use of the Internet in research for both conducting research studies online and studying computer-mediated social behavior. We have also expanded our treatment of conducting and reporting meta-analyses. As with any new edition, this version has also been thoroughly updated to reflect changes and advances in research practices, including new citations to current literature on methodology and procedures.

PART I

Introduction to Social Research Methods



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

1

BASIC CONCEPTS

When American astronauts landed on the moon in the summer of 1971, their activities included an interesting and, for some, surprising demonstration. They showed that when the effects of atmospheric air friction are eliminated, a light object (a feather) and a heavy object (a hammer) will reach the ground at the same time if dropped simultaneously from the same height. This verification of a basic principle of high school physics delighted many viewers of the live televised broadcast, but probably few of them considered the fact that for hundreds of years before Galileo (who is thought to have originally proposed this test), Western scholars had accepted Aristotle's hypothesis that heavy objects would always fall faster than lighter ones. For most of us, Aristotle's assumption seems intuitively correct, even though we now know that it is incorrect and contrary to scientific theory. Not all scientifically demonstrated phenomena contradict "commonsense" intuitions in this way, but this case serves to illustrate the difference between science and intuition as bases of understanding the physical and social world.

Subjecting all theoretical concepts, hypotheses, and expectations to empirical demonstration – that is, testing our ideas and observing the outcomes of these tests – is what distinguishes the scientific method from other forms of inquiry. The **scientific method** is a general approach for acquiring knowledge using *systematic* and *objective* methods to understand a phenomenon. The scientific method provides an overarching methodological blueprint that outlines the steps to be taken to test our ideas.

Objectivity requires that the outcomes of a scientific test be discernible and measurable. The requirement that science deals with observables plays a role in continued scientific advance, as it prods us to continually develop ever more sensitive instruments, methods, and techniques to make observable the formerly invisible. A second requirement is that recorded observational data can be verified by others who either witness the same event or replicate the observational method. The astronaut as well as the television audience observed the same outcome of the feather and hammer study, and therefore likely arrived at a similar conclusion, that objects fall at a rate independent of their mass once air friction is eliminated.

A final requirement is to control for extraneous conditions and variables that might call into question the results of a test. In attempting to release the feather and hammer simultaneously, for example, the astronaut was employing a very rough form of the scientific method. Although probably not discernible to the naked eye, both objects were unlikely released at precisely the

4 Introduction to Social Research Methods

same moment, so a more carefully controlled and systematic study might drop both objects concurrently by a mechanical claw in a vacuum devoid of air resistance.

Outlining the principles of the scientific method, which lend structure to the procedures used to conduct systematic inquiries into human behavior, is what this book is all about. The book is intended to present the research methods and designs that have been derived from basic principles of the scientific method, and to show how these techniques can be applied appropriately and effectively to the study of human cognition, affect, and behavior in social contexts.

Science and Daily Life

It is important to understand that the research principles and techniques presented throughout this text are not reserved solely for the investigation of major scientific theories. At issue in many instances are questions of purely personal interest – the consensus surrounding one’s beliefs, the relative quality of one’s performance, the wisdom of one’s decisions – and in these circumstances, too, the application of the scientific method can prove useful. At first glance, using scientific methods to guide one’s own decision-making processes might appear somewhat extreme; however, considering much current research on human judgment that demonstrates the frailty of our decision-making powers, this is a fitting application of research methods, especially when issues of personal consequence have the potential to affect others as well.

The susceptibility of people’s judgmental processes to a host of biasing influences is well documented (e.g., Kahneman et al., 1982; Sanna & Schwarz, 2006; Van Bavel & Pereira, 2018; West & Kenny, 2011). Research suggests it is risky to depend solely upon one’s own opinions or intuitions to evaluate the truth of an assertion or to make predictions. In judging what people might do under given circumstances, for example, we are inclined to rely heavily, perhaps too heavily, on our own insights, feelings, and interpretations, and to assume that other reasonable people would think, feel, and act just as we do. This is known as the “false consensus” effect (Ross et al., 1977) and is an apparently common and ubiquitous judgmental bias. There is no simple solution to such problems, but there is an available alternative, namely, to test our intuitions, opinions, and decisions systematically, rather than merely to assume that they are valid or commonly accepted.

The specific purpose of this first chapter is to acquaint readers with the fundamentals of research using the scientific method, and to introduce several important themes that run throughout the text. There are many controversial issues in the philosophy of science – such as the status of induction or the logical framework for theory verification (e.g., Bhaskar, 1978; Kuhn, 1970; Manicas & Secord, 1983; Popper, 1959, 1963) – but these more philosophical concerns will be avoided in favor of a more descriptive and explanatory presentation of the “ground rules” of scientific inquiry as agreed to by most social scientists.

From Theory, Concept, or Idea to Operation

The common feature of all approaches making use of the scientific method, regardless of discipline, is the emphasis on the study of observable phenomena. No matter how abstract the generalization or explanatory conceptualization at the theoretical level, the concepts or ideas under investigation must be reduced to, or translated into, observable manifestations. Thus, conducting a scientific study entails a process of converting abstract ideas into a form that can be observed and recorded.

Figure 1.1 represents pictorially this translation of theoretical concepts into research operations. In the first phase of the translation process, the researcher’s general theory, concept, or

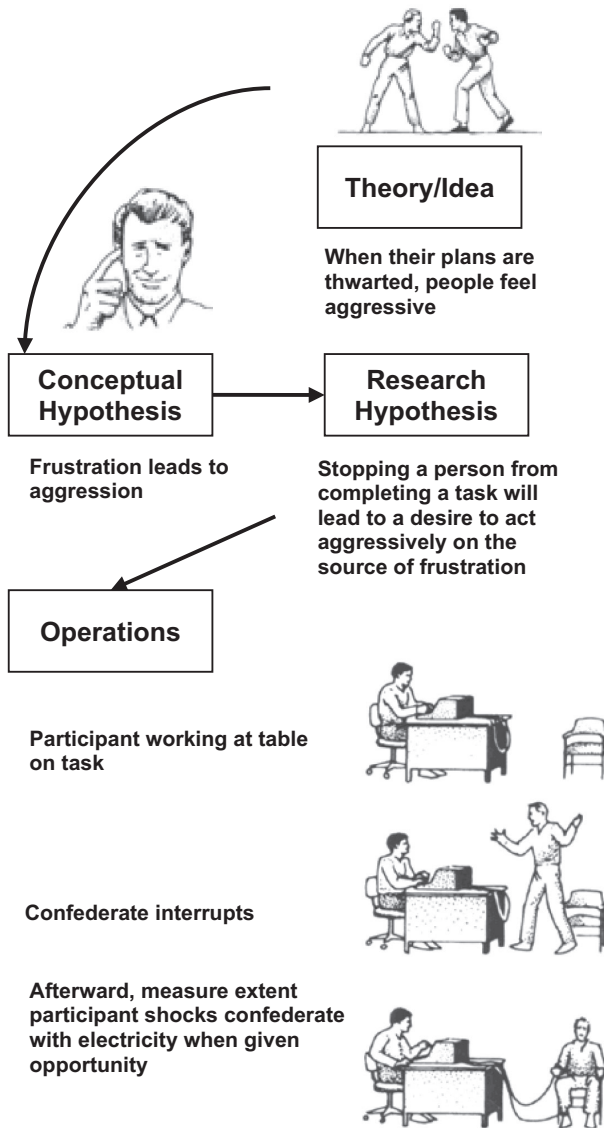


FIGURE 1.1 Translating a theory or idea into research operations.

idea is stated specifically in the form of a **conceptual hypothesis** or prediction about the relationship between two or more theoretical constructs. In this figure, the conceptual hypothesis is about a predicted relationship between a psychological state (frustration) and a subsequent behavior (aggression).

At a finer level of specificity is a **research hypothesis**, which is an empirical specification of the conceptual hypothesis and may be viewed as a prediction about what will be observed in the study. Figure 1.1 illustrates how the conceptual variables of frustration and aggression are empirically specified as being interrupted in an ongoing task and desiring to aggress against the interrupter. A hypothesis is usually a testable *directional* prediction about the specific relationship being studied. In the absence of a theoretically based prediction, the study may be framed in

terms of a **research question**, a nondirectional statement about a specific relationship that ends with a question mark (“Are people with higher pain management skills more likely to survive after major surgery?”). A hypothesis, on the other hand, is expressed as a statement regarding the anticipated direction of the results (“People with higher pain management skills will be more likely to survive after major surgery.”). There are many ways that such hypotheses are formed, and we consider some of these in the following section.

Generating a Hypothesis or Research Question

The development of hypotheses is one of science’s most complex creative processes. As McGuire (1973, 1997) and others (e.g., Jaccard & Jacoby, 2010) have observed, instructors and researchers alike have been reluctant to attempt to teach their students this art, believing it to be so complex as to be beyond instruction. However, by following the lead of some of the field’s most creative researchers, we can learn something about the means that they employ in developing hypotheses.

Theory. The most important, straightforward, and widely used technique of hypothesis generation involves the logical deduction of expectations from some established theory. This is often referred to as the **hypothetico-deductive model** of scientific research. A **theory** consists of a set of premises about why or how things occur or about the processes underlying those relations. A theory serves as an overarching framework for explaining events in a particular domain and for deriving expectations about causal relationships (i.e., what leads to what). The general form of hypothesis deduction from a theory is as follows:

Theory X implies that A results in B

Based on this logic, if the assumptions in theory X are true, a conceptual hypothesis might be generated to test the prediction that producing A will result in the occurrence of B. These hypothetical deductions are based on the tentative premises, assumptions, and implications advanced in the theory. Figure 1.1 illustrates a classic case of the hypothetico-deductive model where a conceptual hypothesis is derived from an abstract theory. The hypothesis that frustration leads to aggression illustrated in the figure was originally derived from psychoanalytic theory of aggressive drive (Dollard et al., 1939). Keep in mind that many hypotheses could be creatively generated to test any single theory. Thus, this very same theory might enable the inference of a hypothesis statement for another study proposing that A also produces C. A theory operates at a relatively abstract level, and therefore is tested only indirectly via the observed data of studies testing research hypotheses generated from the same theory.

Conflicting observations. A second source of generating hypotheses arises from conflicting findings in the published research on a given phenomenon. Typically in this approach, the researcher searches for a condition whose presence or absence helps to explain observed variations or conflicts in findings in an area of investigation. This approach helps to refine theory by testing hypotheses that provide a stricter specification of the conditions under which a particular outcome can be expected to occur (or not to occur).

An example of the use of the “conflicting findings” technique can be seen in research that examined the relationship between ambient temperature and the tendency of people to act aggressively. Experimental research conducted in laboratory settings by Baron and colleagues (e.g., Bell & Baron, 1990) consistently demonstrated that high temperatures *inhibited* the influence of aggressive models. Under normal room temperature, a **confederate**, an actor or accomplice hired by the researcher to pretend to be another participant in the study, could

induce considerable aggression on the part of naïve participants; however, when the ambient temperature of the laboratory was raised, participants' aggressive tendencies diminished. These findings from the laboratory were in marked contrast to those observed by Anderson and his colleagues (e.g., Anderson, 1989; DeWall & Anderson, 2011) outside the laboratory, when they examined the average temperatures of the days on which major riots took place in the United States. They found a positive relationship between these two measures – riots were *more* likely to occur when temperatures were high, suggesting that heat provokes rather than inhibits the spread of aggressive behavior in naturalistic contexts outside the laboratory.

One possible means of reconciling these apparently conflicting results involves an explanation based upon the uniqueness, or prominence, of the temperature in the two research settings. In naturalistic settings (as reviewed by Anderson & DeNeve, 1992), we adjust to the temperature. Although a heat wave is obviously uncomfortable, we experience the discomfort more as a dull ache than as a searing pain. Because this irritation is relatively constant, we do not consciously identify our discomfort as being caused by the heat. Under these conditions, an extraneous event in the environment – such as a confrontation between a policeman and a minor traffic offender – might be misinterpreted as the source of discomfort. Thus, the reactions of a crowd of people are likely to escalate when temperatures are high and have been for some time.

In the case of the laboratory research of Baron and his colleagues, however, the high ambient temperature of the laboratory typically comes on abruptly. Participants walk into the laboratory from a relatively normal environment and find themselves in the middle of a heat wave. Under this circumstance, participants readily identify the source of their discomfort, and this discomfort is unlikely to “transfer” to (or be identified with) other stimuli. This explanation of an apparent contradiction of findings from two different programs of laboratory and naturalistic research gave rise to a new theory, known as “excitation transfer,” which has been used to explain interesting behaviors ranging from violent anger to intense attraction in various contexts (Schippers et al., 2022; Zillman, 1979).

Paradoxes. A third source of hypothesis generation comes from observation of seemingly paradoxical behavior or events. For example, in a classic study, Festinger et al. (1956) found that an extremely reclusive “doomsday” cult became much more publicity conscious and much more active in pushing their beliefs *after* their prophecy concerning the end of the world had been shown to be obviously incorrect. The researchers' attempts to make sense of this apparently paradoxical behavior helped to lay the foundations of Festinger's (1957) theory of cognitive dissonance. This classic theory, originating from preliminary observations, then became the basis for hypothesis generation for countless studies to the present day (see Pearce & Cooper, 2021).

Observation of common events. A fourth method of hypothesis development in social research requires that we attend closely to the common, everyday behavioral tactics that people employ in dealing with others. For example, how can some used-car salespersons promise a car at an impossibly low price, later rescind their offer, and still succeed in selling an automobile that is considerably more expensive than that which was agreed upon originally? Why is it that we are much more likely to agree to a rather major imposition if we have recently “given in” to a much smaller request? Social scientists attuned to issues of this sort have been able to develop some interesting and potentially valuable ideas based on their observations, and to apply these insights to hypothesized topics ranging from inducing charitable donations to AIDS prevention to fostering organ donations (e.g., Burger, 1999; Cialdini, 1988; Siegel et al., 2010).

Case studies. The case study is yet a fifth useful source of hypothesis generation. By concentrating intensively on observing a specific person or event, we sometimes can discern systematic or regular relationships among observations, and these in turn can provide the impetus necessary for developing a testable proposition. Sometimes this approach to developing a hypothesis

is used because insufficient prior research or knowledge exists in this area, and thus a special case of this situation is sought for vigilant scrutiny. These observations are then used to inform later hypotheses to test whether the systematic patterns or regularities that were observed are consistent across other people, situations, and interactions. Some of the most noteworthy examples of the use of the case study in the development of theory are provided by Freud and Piaget, both of whom used this approach extensively (some would say exclusively) in developing their theories. These theories have also guided the development of subsequent hypotheses.

Although this list is far from complete, we hope it provides some idea of the range of possibilities available to the social scientist in developing testable hypotheses (see McGuire, 1973, 1997, or Campbell et al., 1982, for a more complete set of suggestions).

After the hypotheses (or research questions) have been developed, we move to the perhaps less “artistic,” but nonetheless creative phase of operationalization.

Operationalization

From a historical perspective, the social sciences are still quite close to the speculative stages of their initial development. It is only over the last 100 years that we have moved from purely speculative, philosophical explanations of human behavior to a more rigorous scientific approach (see Crano & Lac, 2012). The transition to objectivity in social research was marked by a strong emphasis on **operationalization** – translating an abstract theoretical construct or concept into a concrete specification of procedure and measure, so that it may be observed, recorded, and replicated. The requirement for explicit description of research procedures did much to clarify theoretical issues and to place the social sciences on firmer ground as scientific disciplines.

Operationalization requires that we specify the variables critical in empirically assessing the phenomenon of interest. A **variable** is any characteristic or attribute that may differ and is the basic unit of observation and measurement in a study. Variables may include differences among individuals (e.g., gender, aggressive tendencies, ability to ride a unicycle, level of enthusiasm while reading the textbook), contexts (e.g., room temperature, number of people present), or information (e.g., instruction complexity, message content, task difficulty). Although this textbook largely describes studies involving psychological and behavioral variables among humans, variables may also refer to an array of differences in animals (e.g., level of salivation in dogs) as well as inanimate objects (e.g., type of car driven).

Scientists make use of variables for the purpose of understanding **relations** – how gradations or differences in one variable (or a set of variables) might predict or correspond to gradations or differences in another variable (or a set of variables). For instance, we might want to study whether social influence plays a role in the type of car an individual drives (foreign or domestic) or whether students who spend more time and effort in critically reading this textbook will receive a higher grade in their research methodology course.

Contrast the definition of a variable with that of a **constant**, an attribute on which the value is fixed (or identical) across individuals, and therefore is empirically uninteresting and cannot be used to explain differences that might be observed among them. It would be wasteful of resources, for instance, to conduct a study at many elite boarding schools to test the effects of family wealth on academic achievement. If wealth is uniformly high in the study sample, we cannot use it as a predictor. However, in contexts in which income is not a constant, asking this same question may produce a useful and interesting result.

The translation of conceptual variables to scientifically researchable – and thus observable and measurable – variables generally requires two steps. The first involves the redefinition of the

abstraction of a conceptual variable into an empirical definition; that is, the variable is specified in such a way so that it could be manipulated or observed. Of course, what can be observed is inevitably a function of perceptual skills and available instrumentation – what can be seen with the naked eye is different from what can be observed with the aid of a high-powered electron microscope. Thus, what is “objective” or observable must be defined in terms of the current limitations of our senses and technology. (See Chapter 18 for methods that allow scientists to “read” the implicit content of the human mind – content that previously was not available for scientific investigation.) In general, an observation is considered sufficiently objective if independent researchers with similar training and available technical aids can agree upon its evaluation or assessment.

The second step of concept translation involves a specification of the procedures and instruments required to make the actual observations, detailed sufficiently so that other scientists could duplicate the observation for purposes of replication or validation. This translation is referred to as the **operational definition** of the conceptual variable and requires close attention as the most rigorous aspect of scientific methodology. For purposes of empirical testing, it is essential that very specific and precise delineation of empirical definitions be provided, as suggested in the first step. However, it should be made clear that this specification of research operations is not the end product of the scientific investigation, but merely a necessary step in the research process.

Imperfection of operationalizations. Operationalized constructs are used in testing hypotheses and refining theory. As we emphasized, the operational definition is the result of a specification of the research processes employed in the investigation of a given phenomenon in such a way as to make it directly observable, which lends itself to measurement. One’s measurement instrument, for example, might be labeled an attitude test, but what is being observed and, hence, what constitutes the operationalization of “attitude” are the directly observable pencil marks that an individual has made on the questionnaire we have provided. Most theories in the social sciences, however, are concerned with processes that are somewhat removed from what is directly observed. (We are not especially interested in pencil marking behavior.) Whether attempting to examine anxiety, communication skill, or attitude change, for example, we are dealing with internal cognitive processes that are only indirectly and imperfectly *inferred from* an individual’s observable actions. It thus becomes extremely important to note whether the operationalization has any psychological reality: Do the processes that constitute the operations meaningfully reflect the underlying processes that give rise to the observable responses? Put another way, researchers should be concerned with the degree of overlap between the operation and the internal processes it is purported to represent.

In the social sciences, theoretical concepts are generally of a high level of abstraction, but they must be defined through operations that can be carried out with available technological aids. We move, for example, from the conceptual definition of a construct such as “attitude” (e.g., a consistent internal evaluation of an attitude object) to the more empirical realm of endorsements of positively or negatively evaluative statements regarding the attitude object, and finally to the specification of a set of instructions and verbatim items that provide the actual measure of attitude in a particular research study (see Chapter 16).

Specific operations are always imperfect definitions of any given theoretical construct because the product of any observation is a function of multiple sources – including observer errors, instrumentation errors, environmental and contextual conditions – many of which may be totally unrelated to the conceptual variable of interest. In the physical sciences many of the factors that affect instrument readings have been identified and can be minimized or corrected, but even so, some slight variations in measurements, which do not correspond to variations in

the attribute being measured, do occur. In social science research these unidentified sources of variation often are considerably more profound and even those factors that can be identified can seldom be successfully eliminated or controlled. Campbell (1969) represented this state of affairs well when he observed:

...Measurements involve processes which must be specified in terms of many theoretical parameters. For any specific measurement process, we know on theoretical grounds that it is a joint function of many scientific laws. Thus, we know on scientific grounds that the measurements resulting cannot purely reflect a single parameter of our scientific theory...Let us consider in detail...a single meter, the galvanometer. The amount of needle displacement is certainly a strong function of the electrical current in the attached wire. But, it is also a function of the friction in the needle bearings, of inertia due to the weight of the needle, of fluctuations in the earth's and other magnetic and electrical fields, of the strong and weak nuclear interactions in adjacent matter, of photon bombardment, etc. We know on theoretical grounds that the needle pointer movements cannot be independent of all of these, i.e., that the pointer movements cannot be definitional of a single parameter...Analogously, for a tally mark on a census-taker's protocol indicating family income, or rent paid, or number of children, we know on theoretical grounds that it is only in part a function of the state of the referents of the question. It is also a function of the social interaction of the interview, of the interviewer's appearance, of the respondent's fear of similar strangers, such as bill collectors, welfare investigators, and the law, etc., etc. A manifest anxiety questionnaire response may in part be a function of anxiety, but it is also a function of vocabulary comprehension, or individual and social class differences in the use of euphoric and dysphoric adjectives, or idiosyncratic definitions of key terms frequently repeated, of respondent expectations as to the personal consequences of describing himself sick or well, etc...

(pp. 14–15)

Given these considerations, it is of critical importance that social researchers recognize that their abstract theoretical concepts are never perfectly embodied in any single measured observation.

Multiple operationalization. Because of the imperfect correspondence between conceptual variables and their observable manifestations, we subscribe to defining a variable using multiple operationalization or multiple operational definitions. **Multiple operationalization** recognizes that no single operation or measurement provides enough information to adequately define a concept a theoretical concept, so the construct is measured through several techniques or operations. Ideally, these techniques should be as different as possible *on irrelevant dimensions*. Any single operationalization of a concept is in error to the extent that it does not completely overlap with the internal physical or psychological state it is purported to represent. However, if many diverse measures of a phenomenon are employed, it is probable that they will have non-overlapping sources of error. They will each miss the mark to some extent but will miss it in different ways. Multiple, diverse measures with non-overlapping sources of error allow the researcher to “triangulate” on the concept. This methodological **triangulation** is the logical outcome, and central advantage, of multiple operationalization. Researchers aggregate the data obtained from these multiple operations to create a more representative, but still imperfect, overall composite, index, or factor. The common variations among heterogeneous observations, all of which focus on the same construct, provide the information necessary to adequately identify, or define, the component of interest (Crano, 1981, 2000).

To gain an intuitive feel for multiple operationalization, consider an individual just waking from a night's sleep and wondering exactly what time of day it is. In the pre-digital world, the

usual procedure for assessing time would be to consult a clock, but clocks vary as instruments of time-keeping. Some are based on electrical power impulses, some are battery operated, some run on mechanical spring loading, etc. The interesting thing about these variations in power source for our various timepieces is that the clocks are each subject to different kinds of potential inaccuracy, or error. Electric clocks are affected by power failures or temporary disconnection from a central power source; batteries die; and springs can be over- or under-wound or fatigued by long use. Consequently, consulting any single clock, whatever its type, leaves some room for doubt about the accuracy of its time assessment. When three different clocks all agree on the time, our confidence is increased, although just how confident we can be depends on whether the three timepieces share a common source of inaccuracy or bias. If all three clocks are electrically powered, they could have been simultaneously affected by the same 20-minute power outage during the middle of the night, and hence all would be reporting the same inaccurate reading. When three clocks with three different sources of power all agree, however, our confidence is enhanced considerably. When the three do not agree, the nature of their disagreement and our knowledge of the different types of inaccuracy to which each clock is subject can help us to track down the sources of error and zero-in on the correct time.

Visualizing the process of triangulation via multiple operationalization is helpful. In Figure 1.2, the black square corresponds to the “phantom” construct or concept, for example, the stress experienced by college students in taking an advanced research methods course. Each circle represents a different operational indicator designed to capture underlying stress levels – but what are the conceptual properties that constitute the very abstract idea of being “stressed out”? Imagine that operation 1 is defined using a psychological measure asking you to estimate from 0 to 10, “What is the level of stress you are experiencing?”; operation 2 is a biological measure of cortisol concentration in your blood; and operation 3 is a behavioral measure of the number of times you awaken in the middle of the night in a cold sweat. Diverse operations should be sought, as they may help us measure different features of the construct. As depicted in how much each circle overlaps with the square, a single operation is sorely inadequate in capturing the

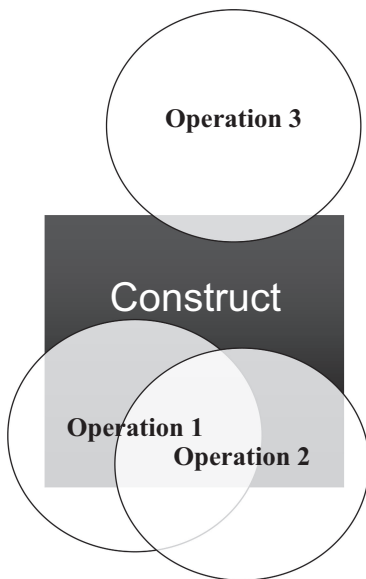


FIGURE 1.2 Triangulating on a construct via multiple operationalism.

conceptual breadth of the construct, but a set of operations increases the likelihood that a greater percentage of the construct will be represented in the study. As might be the case if different quality measures are undertaken, operations 1 and 2 are better manifestations of the underlying construct than operation 3. Also notice that within the construct, operation 1 partially overlaps with operation 2, indicating that portions of these measures tap the construct in the same way.

From Operation to Measurement

The product of most scientific observations is expressed in terms of some scale of *measurement* (Stevens, 1951). A **measurement scale** refers to the assignment of numbers to specific observations in such a way as to reflect variations among those observations. The level of measurement, or the degree of correspondence between number assignment and extent of variation in the attribute being observed, depends upon the rules of the measurement scale assignment. The simplest level of correspondence is represented by the **nominal scale**, a measure requiring only that different numbers be assigned to observations that can be differentiated in some discernible way. Values assigned to nominal scales are arbitrary and therefore not meaningful, as higher and lower values do not signify more or less of that variable; they are merely *categorical*. Examples include coding of participants into qualitative variables representing categories such as gender (male or female or binary), racial or ethnic identification (White, Black, Latino, or Asian), or political partial affiliation (Republican, Democrat, or independent).

Somewhat more complex than the requirements for the application of the nominal rule are those measurements that represent an **ordinal scale**, a measure requiring higher scores to represent greater levels of the construct (the ordering of number labels corresponds to the ordering of observations) and therefore ranked on the construct. Ordinal scales are commonly employed to rank variables along such dimensions as degree of anxiety arousal (high, medium, low), socioeconomic status (upper, upper middle, middle, lower middle, etc.), or finish in a hot chili pepper eating contest (first place, second place, third place, etc.).

Even more measurement sophistication is achieved using an **interval scale**, a measure requiring higher scores to represent greater levels of a construct, and equal distances in numbers reflect equal gradations in different observations. The Fahrenheit scale is an instance of interval scale, as a ten-point difference between 70 and 80 degrees indicates the same change in heat as the difference between 90 and 100 degrees. A Fahrenheit reading of zero, however, does not signify the absence of heat, as this temperature scale may assume negative values. Measures of attitudes derived from pair comparison scaling techniques (Chapter 17) provide examples of interval scales in social research.

A fourth type of measurement scale requires the application of number assignment using the **ratio scale**, a measure requiring higher scores to represent greater levels of a construct, equal distances in numbers reflecting equal gradations in different observations, and a meaningful absolute zero-point. In ratio scaling, the zero-point meaningfully reflects the absence of the attribute being measured. For interval scaling, it is sufficient that the different distances reflect comparative values of differences in observations rather than measurements against some absolute zero-point, but both criteria are required for ratio scales. The requirements of a ratio scale are satisfied in some measurement techniques, such as the amount of time taken to make a response, the amount of pressure exerted on a hand-press device, and the number of inches of rainfall – all with meaningful zero points. A decision tree that illustrates the requirements for different levels of measurement scale is presented in Figure 1.3.

For practical purposes, the essential distinction among these four levels of measurement is the extent of correspondence between the numbers recorded by the researcher and the subtle

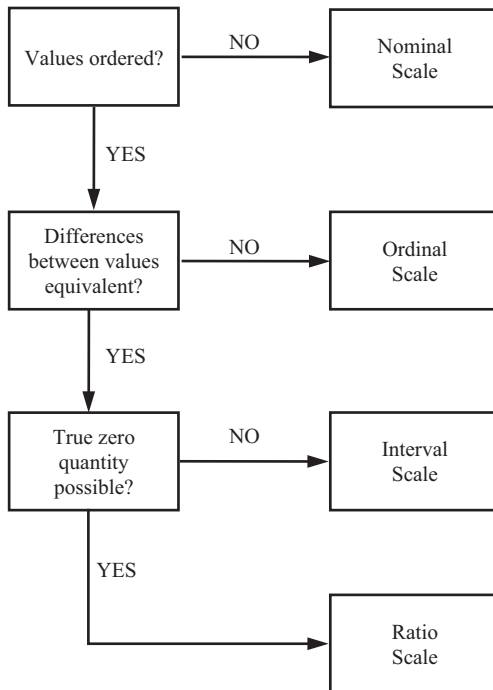


FIGURE 1.3 Decision tree to determine nominal, ordinal, interval, or ratio level of measurement.

variations in the attribute or construct being measured, that is, in the “fit” between assignment of numbers and observations. The higher the level of measurement, the greater the number of gradations in the data are represented in the numbers given by the researcher. We do not always aim for the highest level of measurement, however, because they make increasingly greater demands on data quality, and on respondents’ abilities to successfully differentiate the values along the scale. Higher levels of measurement make the variable more sensitive in detecting subtle nuances in observations and responses. Thus, the ability to interpret observations in terms of higher-order measurements adds considerable precision to the research process, but the assignment of numbers does not automatically create precision in measurement if the rules of number assignment have been violated.

Role of Theory in Scientific Inquiry

It is traditional to think of scientific theories as convenient data-summarizing generalizations that are helpful in guiding decisions about the content and interpretation of future observations, but that are discarded when obtained observations do not fit the conceptual framework. The actual role of theory in most scientific endeavors, however, differs from this ideal. In our view, a formal theory is a template, or a pattern, against which various data sets, obtained from studies testing hypotheses, are compared. The extent of the “match” between theory pattern and data pattern provides an indication of the usefulness of the theory. Of course, the match is rarely, if ever, perfect, but with the continual modification of the theoretical template, an increasingly better fit can be obtained. Even when the fit between data and theory is unsatisfactory, the theory is rarely discarded until an alternative theory that provides a better match with the data is available.

As suggested in the preceding paragraph, the interplay between theory and data is not entirely an “objective” process. Kuhn’s (1970) still stimulating and controversial analysis of scientific “revolutions” has produced increased awareness of the influence of social factors on the process of scientific advance (Bird, 2000; Horwich, 1993). The acceptance of theoretical positions is partly determined by the prevailing social climate, and frequently in large part by the personalities of the advocates of competing theoretical perspectives. Furthermore, vigorous and valuable research is not necessarily inspired by formal theoretical considerations. Some active areas of research have been derived from essentially exploratory investigations, inspired more by hunches and guesswork than by formal theory.

Ideally, the theory-testing process begins with the derivation (from the theoretical structure) of implications that can be stated in the form of hypotheses regarding the existence of relationships among observable variables. Most of these derived hypotheses are of the general form, “if theory U is true, then, with irrelevant factors eliminated, the occurrence of variable A should be related to the occurrence (or condition) of variable B.” Comparisons between these predicted relationships and the actual outcomes of controlled observations comprise the testing process, through which the original theory is subjected to continuous potential disconfirmation (although the failure to obtain the predicted results *in any particular case* can always be attributed to the inadequacy of the empirical statement or the research operations to reflect the theoretical position). Successive failures of empirical investigations to *disconfirm* the derived hypotheses create incremental acceptance of the theoretical position. This process continues until some outcome is encountered that cannot be explained by the current theory at a time when some alternative theory accounts for all the previous findings *and* the otherwise inexplicable result.

There are, of course, degrees of uncertainty. A scientist whose theory has been tested only once or twice should be more concerned about its validity than one whose theory has been shown to be consistently successful in the face of strong, plausible challenges. Thus, the general approach emphasized throughout this text will consist of various strategies whose major function is the *falsification* of as many competing hypotheses as possible. Through this continuous process, the gradual purification of theory can be accomplished, though certainty is never assured. Again, however, it must be stressed that even long-term success and widespread acceptance is no guarantee that a theory is true or valid or even optimally useful.

In the case of scientific theory, a position rarely stands or falls because of a single observation. Every individual study is subject to several potential explanations. Only the total pattern of multiple test outcomes determines the goodness of fit between theory and data. Thus, the process of theory testing via the measurement of observations is best viewed in terms of multiple operationalization across many studies. A theory frequently stands in the absence of perfect fit between its implications and every data outcome *until* some alternative theory that accounts for more of the actual observations becomes available. The more convergence there is between the hypothesized outcomes of successive heterogeneous investigations and the implications of a particular theoretical position, the less likely it becomes that some alternative explanation will become available to replace it, although the possibility of such an alternative can never be ruled out entirely.

Conclusion and Overview

The preceding section on the role of theory in social research presents a view of research as an iterative, cumulative process extended across time and persons. Any single research study should be regarded as a part of this larger scientific enterprise—much like an individual piece in a giant

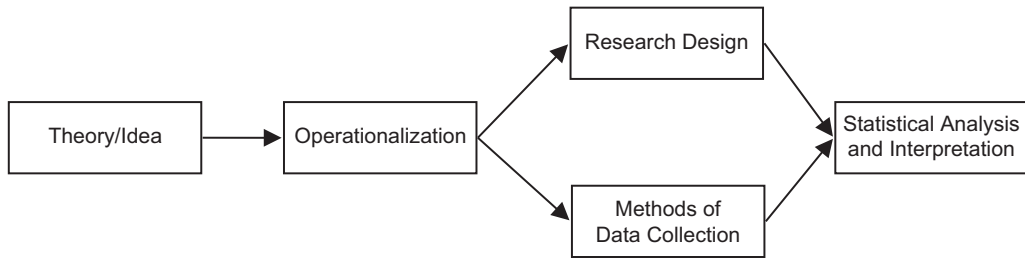


FIGURE 1.4 Stages of research.

jigsaw puzzle. No single study will be sufficient in and of itself to prove or disprove a theory or hypothesis (Greenwald, 1975), but each study contributes in some unique way to piece together the total picture. For the individual researcher, however, the major focus of attention will be the design and conduct of one research project at a time. It is the purpose of this book to introduce students of social research to the basic tools needed for that purpose.

The stages of any individual research study are diagrammed in Figure 1.4. We have already discussed, briefly, the principles involved in translating conceptual variables into empirical definitions and then operational definitions. The discussion of research ideas further may be divided into two basic steps. The first is the selection of an overall research design – the master plan that defines the systematic procedures we are conducting. Research design includes decisions as to the setting in which the research is to be undertaken, the relevant variables to be included, and whether the researcher controls or manipulates critical features of the situation. The second step is that of choosing the specific methods of data collection to be used in the study. How are the selected variables of interest to be assessed, and what are the respective roles of the researcher and the participants in the data collection process?

The combination of research design and data collection methods helps to define the research study. When the design and methods have been carried out, the final stages of the study involve analysis and interpretation of the resulting data. Correct interpretation of the results of such analyses must consider the design and methods by which the data were obtained, and these are the features of the research process that we will focus on throughout this text. It is of utmost importance to craft a well-designed study with methods appropriate to the topic of investigation, as failing to do so will produce a host of problems that subsequently compromise observed outcomes, making these results susceptible to interpretation problems that cannot be remedied by statistical analyses.

Chapters 2 to 4 are devoted to general issues in the operationalization of theoretical concepts and the selection of research design and methods of measurement. Chapters 5 to 11 present more detailed discussion of specific research designs that may be employed, elaborating on a basic distinction between experimental and nonexperimental research approaches. Chapters 12 to 19 discuss procedures of specific data collection methods, with attention to the relative strengths and weaknesses of each. Specific issues of analysis and interpretation are raised as they are relevant to materials covered in these two sections. Finally, Chapters 20 and 21 address more general issues that cut across all methods of research in the social sciences. These general issues include the social and ethical responsibilities associated with the conduct of social research and advancing the cumulative knowledge base through combining and synthesizing findings from different primary studies. In these final chapters we hope to make clear that science is a collective enterprise undertaken in the context of societal values.

References

- Anderson, C. A. (1989). Temperature and aggression: Ubiquitous effects of heat on occurrence of human violence. *Psychological Bulletin*, 106, 74–96.
- Anderson, C. A., & De Neve, K. M. (1992). Temperature, aggression, and the negative affect escape model. *Psychological Bulletin*, 111, 347–351.
- Bell, P. A., & Baron, R. A. (1990). Affect and aggression. In B. S. Moore & A. M. Isen (Eds.), *Affect and social behavior* (pp. 64–88). Cambridge: Cambridge University Press, and Paris: Editions de la Maison des Sciences de l'Homme.
- Bhaskar, R. (1978). On the possibility of social scientific knowledge and the limits of behaviorism. *Journal for the Theory of Social Behavior*, 8, 1–28.
- Bird, A. (2000). *Thomas Kuhn*. Chesham; Acumen, and Princeton: Princeton University Press.
- Burger, G. (1999). The foot-in-the-door compliance procedure: A multiple-process analysis and review. *Personality and Social Psychology Review*, 3, 303–325.
- Campbell, D. T. (1969). Definitional versus multiple operationism. *Et al.*, 2, 14–17.
- Campbell, J. P., Daft, R. L., & Hulin, C. L. (1982). *What to study: Generating and developing research questions*. Beverly Hills, CA: Sage.
- Cialdini, R. B. (1988). *Influence: Science and practice* (2nd ed.). Glenview, IL: Scott, Foresman.
- Crano, W. D. (1981). Triangulation and cross-cultural research. In M. B. Brewer & B. E. Collins (Eds.), *Scientific inquiry and the social sciences: A volume in honor of Donald T. Campbell* (pp. 317–344). San Francisco, CA: Jossey-Bass.
- Crano, W. D. (2000). The multitrait-multimethod matrix as synopsis and recapitulation of Campbell's views on the proper conduct of social inquiry. In L. Bickman (Ed.), *Research design: Donald Campbell's legacy* (Chapter 3, pp. 37–61). Beverly Hills, CA: Sage.
- Crano, W. D., & Lac, A. (2012). The evolution of research methodologies in (social) psychology. In A. Kruglanski & W. Stroebe (Eds.), *Handbook of the history of social psychology* (pp. 159–174). New York: Psychology Press.
- DeWall, C. N., & Anderson, C. A. (2011). The general aggression model. In P. R. Shaver & M. Mikulincer (Eds.), *Human aggression and violence: Causes, manifestations, and consequences* (pp. 15–33). Washington, DC: American Psychological Association.
- Dollard, J., Miller, N. E., Doob, L. W., Mowrer, O. H., & Sears, R. R. (1939). *Frustration and aggression*. New Haven, CT: Yale University Press.
- Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: University of California Press.
- Festinger, L., Riecken, H. W., & Schachter, S. (1956). *When prophecy fails*. Minneapolis: University of Minnesota Press.
- Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, 82, 1–20.
- Horwich, P. (Ed.) (1993). *World changes: Thomas Kuhn and the nature of science*. Cambridge, MA: MIT Press.
- Jaccard, J., & Jacoby, J. (2010). *Theory construction and model-building skills: A practical guide for social scientists*. New York: Guilford Press.
- Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.
- Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press. (1st ed., 1962).
- Manicas, P. T., & Secord, P. F. (1983). Implications for psychology of the new philosophy of science. *American Psychologist*, 38, 399–413.
- McGuire, W. J. (1973). The Yin and Yang of progress in social psychology. *Journal of Personality and Social Psychology*, 26, 446–456.
- McGuire, W. J. (1997). Creative hypothesis generating in psychology: Some useful heuristics. *Annual Review of Psychology*, 48, 1–30.
- Pearce, L., & Cooper, J. (2021). Fostering COVID-19 safe behaviors using cognitive dissonance. *Basic and Applied Social Psychology*, 43(5), 267–282.
- Popper, K. R. (1959). *The logic of scientific discovery*. New York: Basic Books.
- Popper, K. R. (1963). *Conjectures and refutations: The growth of scientific knowledge*. London: Routledge.

- Ross, L., Greene, D., & House, P. (1977). The “false consensus” effect: An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, 13, 279–301.
- Sanna, L. J., & Schwarz, N. (2006). Metacognitive experiences and human judgment: The case of hindsight bias and its debiasing. *Current Directions in Psychological Science*, 15, 172–176.
- Schippers, E. E., Smid, W. J., Both, S., & Smit, J. H. (2022). Excitation transfer between sexual arousal and other emotions in a community sample. *Archives of Sexual Behavior*, 51, 3905–3918.
- Siegel, J. T., Alvaro, E. M., Crano, W. D., Gonzalez, A. V., Tang, J. C., & Jones, S. P. (2010). Passive-positive organ donor registration behavior: A mixed method assessment of the IIFF model. *Psychology, Health & Medicine*, 15, 198–209.
- Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 28–42). New York: Wiley.
- Van Bavel, J. J., & Pereira, A. (2018). The partisan brain: An identity-based model of political belief. *Trends in Cognitive Sciences*, 22, 213–224.
- West, T. V., & Kenny, D. A. (2011). The truth and bias model of judgment. *Psychological Review*, 118, 357–378.
- Zillman, D. (1979). *Hostility and aggression*. Hillsdale, NJ: Erlbaum.

2

INTERNAL AND EXTERNAL VALIDITY

A wide range of potential research designs will be covered in the chapters that follow. They can be categorized roughly into either experimental or nonexperimental strategies. **Experimental methods** involve research in which participant actions are limited or in some way constrained by the controlled manipulation of variables determined by the researcher. Ideally, experiments should be characterized by random assignment, a procedure used to place participants, by chance alone, into the different conditions, which enables strong inferences of causality. **Quasi-experimental methods** are a variant of experimental methods lacking in random assignment, but as in experiments, participants are exposed to some form of variable manipulation imposed by the researcher. For example, we might be interested in the effects of National Merit Scholarships on college performance. Obviously, a researcher cannot randomly assign high school seniors to receive or not receive a Merit scholarship, because of ethical concerns and the fact that receipt of the award is determined by scores on a test. Quasi-experimental methods (Chapter 9) are an attempt to make inferences about possible variables responsible for the outcome differences between recipients and non-recipients of the scholarship, and to disentangle the effects of the scholarship from other variables such as intelligence or work ethic.

In **nonexperimental methods**, no variables are manipulated; the relations among naturally occurring variables are measured by the researcher. Generally, nonexperimental studies are characterized by a focus on associations or correlations among concepts that the researcher is unable to manipulate, as they involve variables over which the researcher has little, if any, control – sex, race, religiosity, social status, political party affiliation, intelligent design beliefs, and so on. In nonexperimental studies, correlational techniques could be used to reveal the strength of association among nonmanipulated variables, as the researcher does not interfere with respondents' characteristics and normal behavior. In experimental studies, the researcher is an “active observer,” systematically manipulating variables and observing the outcome of these interventions. In nonexperimental studies, however, the researcher can be characterized as a “passive observer” of naturally occurring variables. In nonexperimental studies, the researcher does not actively intrude on the environment of the investigation. Rather, the associations of naturally occurring variations, or of individual characteristics the participants bring to the research, are the focus of study.

These distinctions between experimental and nonexperimental research are consistent with those advanced by Campbell and Stanley (1963) in their classic discussion of research design. However, these different research design strategies should not be regarded as discrete and mutually exclusive, but rather as points on a continuum about the extent that the researcher has controlled or constrained the context of participant experience, and the manner in which their reports of this experience are restricted by the study arrangements.

The advantages and limitations of experimental and nonexperimental research strategies tend to be complementary, so effective programs of research should make use of both in different stages of the research process.¹ Experiments are deemed the “gold standard” of research because of their methodological rigor and their consequent capacity to foster causal inferences. Experimental studies are usually undertaken in the laboratory because the laboratory provides the conditions allowing the researcher to carefully regulate the context in which controlled variations are implemented and their effects observed. However, this control and precision come at a cost. In research on human behavior, the very control that marks the advantage of experimental techniques places limitations on the representativeness of the phenomena they are used to study, and the generalizations that may be advanced.

Nonexperimental (or correlational) strategies, on the other hand, are more likely to have the value of “real-world” context and the availability of mass data in developing information about human actions. However, these advantages are bought at the cost of a lack of control over nonsystematic variation in the variables of interest. The inability to exert control over critical variables can result in interesting, but scientifically inconclusive findings. The relative value of experimental versus nonexperimental research methods depends to a large extent on the importance placed by the researcher on being able to make causal inferences about the relations among the variables being studied (Brewer, 2000).

Nonexperimental methods are typically undertaken in field research settings, but sometimes are employed in laboratory settings as well. For example, if we invite both male and female participants to come to a laboratory to complete questionnaires regarding their need for affiliation and then assess the difference between the men and women in their questionnaire responses, this is still a nonexperimental (correlational) study. Both participant sex and level of affiliation need are individual difference variables, not systematically controlled or manipulated by the researcher. The fact that the measures were obtained in a lab and that there are two groups being compared does not make this an experiment or quasi-experiment.

Causation

The purpose of many research studies is to investigate a hypothesized relation between the occurrence of variation or change in one variable, A, and the occurrence of variation or change in another variable, B. Variables may be states of the physical or social environment (e.g., weather conditions, the number of people present in the situation), properties of a stimulus (e.g., the facial expression in a photograph, the content of a message), or characteristics of a person or a person’s behavior (e.g., mood state, degree of aggression), to name a few possibilities. Hypothesized and observed relations can involve two environmental variables (e.g., the relation between variations in barometric pressure and rain), between an environmental or stimulus variable and an individual characteristic or trait (e.g., the relation between the state of the weather and the average mood of people exposed to it), or between characteristics of individuals (e.g., the relation between mood and aggressiveness). To say that there is a relation between two such variables means that if the value or state of one variable differs or changes, we can expect that

the value or state of the other will also change or differ. So, for example, if we measure people's mood on a sunny day and then again on a cloudy day, with the result showing that mood is more negative on the second occasion, we can say that the data suggest a relation between the state of the weather and people's mood.

The more precise the theoretical specification of a predicted relation, the more closely the obtained data can be matched against the prediction. The nature of the relation may be specified in terms of the *form* it will take, that is, what kind of changes in variable A will accompany specified changes in variable B, and what the *causal direction* of the relation will be. Directionality of relations may be classified into three types:

- 1 **Unidirectional causation**, a relation in which changes in A produce subsequent changes in B, but changes in B do not influence A (e.g., increases in the temperature-humidity index produces an increase in aggressive responses of rats, but the degree of aggressiveness of rats does not affect weather conditions).
- 2 **Bidirectional causation**, a relation in which changes in A produce changes in B and, in addition, changes in B produces changes in A (e.g., perceiving threat produces feelings of anxiety, and increasing anxiety enhances the perception of threat).
- 3 **Noncausal covariation** (or correlation, or third-variable causation), a relation in which changes in A are indirectly accompanied by changes in B, because both A and B are determined by changes in another variable, C. For example, increases or decreases in the cost of living, C, result in the rise or fall of birth rate, A, and consumption of beef steak, B.

Causation versus Covariation

The simple observation that when A varies, B also varies is not sufficient to demonstrate which of these three possibilities is correct. To determine that any causal direction exists, alternative explanations for the observed relation must be ruled out. Both types 1 (unidirectional causation) and 2 (bidirectional causation) must be distinguished from type 3 (noncausal relation) by demonstrating that when A (or B) changes in isolation from changes in any other variable (C), subsequent changes are observed in B (or A). When an observed relation is found to have been mistakenly interpreted as a causal relation of type 1 or 2, when it is actually a case of type 3, the relation is said to be *spurious*.

Type 1, unidirectional causation, can be distinguished from type 2, bidirectional, by observing whether modifications in A produce changes in B, and vice versa. Finally, the validity of the predicted third variable in a case of type 3 can be determined by noting whether the relation between A and B can be eliminated if C (and only C) is held constant (that is, not permitted to vary). All these differentiations are possible only under conditions in which variations in the relevant variables can be observed uncontaminated by related variations that are not relevant to the theoretical relation.

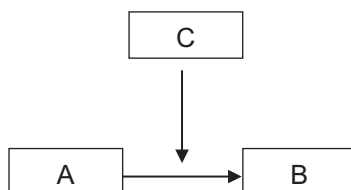
Moderators and Mediators of Relations

In addition to specifying the nature and direction of a relation under study, it is also important to distinguish between two different types of "third variables" that can influence predictive relations – moderators and mediators (Baron & Kenny, 1986; Hayes, 2013). A **moderator** is a third variable that can either augment or block the presence of a predictive relation. To take another weather-related illustration, consider the relation between exposure to sun and sunburn. Although there is a well-established cause-effect link here, it can be moderated by several

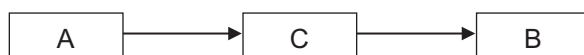
other variables. For instance, the sun to sunburn relation is stronger for fair-skinned than for dark-skinned persons. Thus, fair skin is a moderator variable that enhances the relation between sun exposure and burning. However, this does not indicate that the sun to sunburn relation is spurious. The moderator variable (skin pigmentation) does not produce the effect in the absence of the predictor (sun exposure).

Other moderator variables can reduce or block a causal sequence. For instance, the use of effective sunscreen lotions literally “blocks” (or at least retards) the link between the sun’s ultra-violet rays and burning. Thus, a researcher who assesses the correlation between sun exposure and sunburn among a sample of fair-skinned people who never venture outdoors without a thick coat of 90 SPF sunblock would be ill-advised to conclude that there is a complete absence of the relation across all groups.

Moderator effects can be represented schematically as follows:



Like Baron and Kenny, we think it is important here to distinguish between third variables that serve as *moderators* (as in the illustration) and those that serve as mediators of a relation. With moderator effects, there is a link from A to B, but the observed relation between these two variables is qualified by levels of moderator variable “C” that either enhances or blocks the causal process. On the other hand, a **mediator** is a third variable that serves as intermediary and helps explicate the chain of causation of a predictive relation. A mediational relation is represented as follows:



In this case, the presence of “C” is necessary to complete the directional process that links A to B. In effect, varying A results in variations in C, which in turn results in changes in B. To return to our weather examples, the effect of rain on depression may be mediated by social factors. Rain causes people to stay indoors or to hide behind big umbrellas, hence reducing social contact. Social isolation, in turn, may produce depression. However, rain may not be the only determinant of social isolation. In this case, rain as a predictor variable is a sufficient, but not necessary cause in its link to depression. To demonstrate that A results in B only if C occurs does not invalidate the claim that A and B have a relation; it only helps to explicate the relation in terms of a chain of causation.

Moderator variables are determined by testing the interaction of a predictor variable and moderator variable on an outcome, as will be discussed more fully in Chapter 5. Mediational hypotheses will be covered in Chapters 10 and 11. The role of moderating and mediating variables in predictive relations can be evaluated for both experimental and nonexperimental research strategies.

Causation and Experiments

Although all scientific methods are empirical, not all are experimental. Some areas of investigation – for example, astronomy – involve phenomena that cannot be brought under the

manipulative control of the researcher and can only be observed naturally using nonexperimental strategies. However, where experimentation is possible, it is the most powerful research strategy available for determining the source and direction of relations between variables.

Essentially, the purpose of the hypothesis-testing experiment is to clarify the relation between two (or more) variables by bringing the variation of at least one of the elements in the relation under the control of the researcher; in experiments, that is, the researcher determines when and how changes in this variable occur.

The **independent variable** is a characteristic manipulated independently of its natural sources of covariation to produce different conditions in an experimental study. We refer to conditions or variations of the independent variable as experimental *treatments*, *manipulations*, or *levels*. The experimenter manipulates or controls participants' exposure to one of the variations in the independent variable in some systematic way. This is done to ensure that effects from other natural sources of covariation (or other variables related to this predictor) would not obscure the impact on the outcome of interest.

The **dependent variable** is the measured outcome or consequence, not manipulated by the researcher, and expected to be influenced by (or dependent upon) manipulation of the independent variable in an experimental study. If the experiment is adequately controlled (see Chapters 5 and 6), the observed variations in scores of the dependent variable will be attributable to the effects of variations of the independent variable. Thus, the experiment provides a critical setting for demonstrating the nature of the A to B predictive relation between theoretically relevant variables, but, as we discussed in Chapter 1, it is only one phase in the theory-building process.²

For the most technically correct usage, the terms *independent variable* and *dependent variable* are typically reserved for experimental methods. Many researchers and methodologists, out of convenience, have adopted this same terminology for nonexperimental strategies as well, although this is not entirely consistent with the original intended meaning of these terms. The terms *predictor variable* and *criterion variable* (or outcome) provide more encompassing language to describe the A to B relation that can be used in both experimental and nonexperimental research designs. An independent variable represents a specific type of a predictor variable, and the dependent variable is a special case of a criterion variable or outcome.

Distinguishing Internal and External Validity

The design and conduct of any hypothesis-testing experiment should be guided by considerations of two types of validity – internal and external validity (Campbell & Stanley, 1963). These forms of validity reflect upon the quality of different aspects of the research process and the conclusions that we wish to draw from the results of a study. **Internal validity** is the extent that inferences of causation can be made about the obtained relation between the independent variable and the dependent variable. Confidence in causal inferences is enhanced when the treatment or manipulation is under the systematic control of the researcher. **External validity** refers to generalizability, the extent to which observed results can be applied to other respondent groups, different settings, and different ways of operationalizing the conceptual variables.

Typically, discussion of internal validity is reserved for research that involves experimental methods, and this is proper, because the issue of internal validity is concerned with the appropriate interpretation of the causal relation between an independent and a dependent variable, the central feature of all experimentation. Considerations of internal validity, however, also may be applied when evaluating relations observed in nonexperimental contexts, but because such

studies do not manipulate variables, causal inferences about their naturally observed relations generally are regarded as being lower in internal validity.

Considerations of external validity or generalizability of results are equally important in evaluating the worth of experimental and nonexperimental research, and as such should be considered in all types of research. Internal validity and external validity, as well as other types of validities, should not be evaluated in black and white terms – whether a study possesses one or the other or both – but rather as points on a continuum. Before discussing the internal-external validity distinction further, however, we will consider three closely related issues – the role of statistics, the field/laboratory distinction, and the distinction between basic and applied research.

Role of Statistics (Statistical Conclusion Validity)

Just as the choice of research method must depend upon considerations of the nature of the phenomenon of interest, so too must the role of statistical techniques be evaluated with respect to the general goal of eliminating or reducing the plausibility of rival alternative hypotheses for the events under investigation. One potential rival explanation that plagues social research at all stages of investigation is *chance*. The phenomena of interest in the social sciences are generally subject to considerable nonsystematic variation – variations from individual to individual, for example, and within individuals from one time to another. Given such uncontrolled (and unexplained) variability among observations, the occurrence of any observed relation or effect is always potentially attributable to the operation of chance, or random happenstance. Statistical analyses assess the possibility that chance is a reasonable alternative explanation of any relational finding.

When investigating a phenomenon, studies almost always collect data from a limited sample of participants (Hallahan & Rosenthal, 1996). We do so because it would be prohibitively difficult and expensive to collect data from all individuals in a target population. Thus, a study's sample is always susceptible to *sampling error*, because the outcome observed in a particular sample, by chance, is not perfectly consistent with what would be found if the entire population was used. This uncertainty due to sampling error is partly responsible for the heavy use of probabilistic language and triple negatives (e.g., “The failure to reject the null hypothesis...”) that serves to terrify so many students of elementary statistics.

The purpose of inferential statistics (e.g., t-tests, analyses of variance, multiple regression) is to assess the likelihood that an obtained effect could have occurred by chance in the absence of any systematic relation between the variables being studied. Statistical inference allows us to assign a probability of the operation of chance due to sampling error as a possible explanation of any relation discovered. Results of a statistical inference test tell us the probability of a **Type I error** of inference – the probability that the observed effect would have been obtained, by chance, if the **null hypothesis** (i.e., no true relation between the predictor and outcome variables) was actually true.

Statistical significance is achieved if the probability of obtaining the observed effect, by chance, is so low as to render the chance explanation implausible. Usually, a Type I probability value of .05 or less (i.e., an outcome that could have occurred by chance no more than five times in 100) is chosen as the cutoff value in determining whether or not the observed relation is defined as statistically significant.³ Suppose you thought that high school boys' participation in contact sports (football, hockey, etc.) was associated with their scoring higher on a test of aggressiveness, and that the more sports they participated in, the higher would be their aggressiveness scores. You gather data on extent of boys' participation, administer a good test of

aggressiveness, and find a statistically significant association (at $p < .05$) between predictor and criterion variables. You are happy that your hypothesis was supported, but remember, there is a 5% chance that you might have obtained this result by chance, even if there was no actual relation. If 100 researchers replicated your study, it is possible that five of them would find results such as yours – by chance. Tests of statistical significance provide insight into the likelihood that chance is a viable alternative explanation of our findings.

When the probability of a Type I error is not low enough (e.g., not below $p < .05$) to rule out the null hypothesis, we should be concerned with a **Type II error** of inference – the probability of failing to reject the null hypothesis when it is actually false (i.e., there really is a relation between the predictor and outcome variables, but our study failed to detect it at above chance level). Reducing the probability of Type II error requires that we design studies with sufficient power (see Cohen, 2003) to detect an effect above and beyond random variation. **Statistical power** is the probability of obtaining a statistically significant effect, if indeed that effect truly exists.

It is desirable to make sure that your study has high statistical power, because if it does not, valuable resources may be wasted in conducting a study that fails to obtain adequate evidence for statistical significance of results. The power of a study depends in part on a number of properties of the study (Cohen, 2003): (a) number of participants (power increases as the number of participants in the study increases), (b) the reliability of the measures that are used (power increases when reliable measures are used, see Chapter 3), (c) the strength (effect size) of the relation, because as the strength of the observed effect increases, so does the power to detect it, and (d) the Type I critical value that is used for the statistical test (tests using a critical rejection region of $p < .05$ have more power than those using a lower probability of $p < .001$, for example). Ideally, our studies should be designed to have sufficient statistical power to attain statistical significance, allowing us to identify real relations when they exist.

Figure 2.1 describes the four possible outcomes that can be found when testing whether a study's outcome corresponds with reality. To give an applied example using this figure, consider you taking a home self-test for the presence of Covid during the coronavirus pandemic. We know that four outcomes are possible in this example: (a) the test is positive and indicates you have Covid and you actually do (correct diagnosis); (b) the test indicates you have Covid but you actually do not (Type I error); (c) the test is negative and indicates that you do not have Covid when you actually do have the virus (Type II error); (d) the test indicates that you do not have Covid and you actually do not (correct decision).

Statistical analyses are powerful and necessary research tools. However, the time and effort required to master the theory and calculations essential for the proper use of statistical

		Actuality/Reality	
		Effect	No Effect
Test/ Outcome	Effect	Correct Decision (Power)	Type I Error
	No Effect	Type II Error	Correct Decision

FIGURE 2.1 Possible scenarios involving a test result and correspondence with reality.

significance testing techniques have led many students of social science to confuse statistical sophistication with expertise in research design. Statistical considerations are not the beginning and end of our research design concerns. Indeed, in our view, proper research design almost invariably simplifies the statistical analyses that we require (Smith, 2000). The statistical analyses that are chosen should depend upon their relevance to the theoretical issues being addressed and the nature of the design that is used. *Complicated statistics cannot compensate for poor design.* Learning to avoid the pitfall of equating current tools of analysis with the *purpose* of research methodology is a major aim of this text. Our sentiments lie with those of Agnew and Pike (1969, p. 142) who observed:

...The researcher who works outside the laboratory should, if anything, be more sophisticated about the principles of measurement, research design, statistics, and rules of evidence than the laboratory researcher. Note that we are talking about principles of research design, not about rituals. Though more difficult to teach, it would be preferable to provide the student with an appreciation of the main rules of evidence used in science. Hopefully, the student would then be able to select research tools that fit his research interests, rather than looking for a research problem that will fit his research tools.

Field versus Laboratory Research

For a long time in social research, a controversy existed between the proponents of field research and those who favored laboratory studies. Field researchers claimed that only in real-life settings could we discover anything of value – that the responses of participants who were studied in the cold, antiseptic environment of the laboratory could not be viewed as valid representations of the behavior they would emit in more normal, everyday, circumstances. In our terminology, they were arguing that laboratory studies have no external validity (generalizability). Laboratory researchers, on the other hand, argued that so many theoretically extraneous events occurred in the natural environment – the field – that one could never be certain about the true relation that existed between independent and dependent variables.

As is the case with most arguments of this type, both sides were partly correct – and both were partly wrong. To be sure, there are numerous examples of laboratory studies that are so devoid of reality that their practical or scientific utility must be seriously questioned. It is just as obvious, however, that not all laboratory research is psychologically, behaviorally, and socially “unreal” (Aronson et al., 1998). In many areas of investigation – communication, persuasion, group interaction – we have gained much valuable information about the complicated nature of human social interaction. To argue that nothing of generalizable value can come out of the social laboratory is to deny the obvious.

On the other side of the argument are the laboratory researchers who contend that so many uncontrolled events occur in the relatively free-form field environment that no clear specification of relations between or among theoretically implicated measures is possible. As before, there are plenty of examples that would serve to “prove” the contentions of the critics of field research. The inability to control for powerful factors that could have major influences on critical behaviors is one of the more telling problems mentioned by critics. The problems introduced by this lack of control are real and are part of the standard set of difficulties field researchers face whenever studies of this type are mounted. With the increasing sophistication evident in much contemporary field research, however, it is becoming ever more apparent that these difficulties can be surmounted – or, if not completely offset, their effects at least identified, as we will discuss in Chapter 8.

There is a place in the social sciences for both field and laboratory research. Each reinforces the value of the other. The findings of the laboratory are retested in the field, where their robustness is put to a severe test. If the findings hold, then we may gain a potentially important addition to our knowledge base. Likewise, the less-than-completely-controlled observations of the field researcher can be brought to the laboratory for more rigorous examination. If these observations prove valid within the stricter confines of the laboratory, their value is already established, given their initial development in the “real world.” Thus, field and laboratory research do not compete, but rather complement each other in real and important ways. Many of today’s leading social researchers agree with this assessment. Many internally valid research endeavors are, in fact, field experiments (Chapter 8); and some of our most powerful generalizations were born in the sterile confines of the laboratory (Mook, 1983). The integration of field and laboratory research methodologies using experimental and nonexperimental strategies is examined in the forthcoming chapters.

Basic versus Applied Research

Viewing the research process as the accumulative reduction of plausible alternatives to a particular theoretical account provides a context for considering the traditional distinction between “basic” and “applied” research. Essentially, the difference between the two lies in whether relatively long-term or short-term gains are expected from the outcomes of the research. The “applied” label refers to those research efforts that are directed toward achieving a specified outcome (e.g., which of several advertising campaigns will produce the greater number of product sales; which serum formula will terminate the symptoms of skin cancer most effectively; which remedial program will reduce illiteracy in the urban ghetto). Because the goals of applied research are relatively concrete, we can determine whether the intended outcome is successfully obtained or not rather quickly. For basic research, the goal of a research project is to contribute to that ephemeral universe of knowledge or, in more specific terms, to add to the accumulative pattern of data that will ultimately determine the survival value of alternative theoretical interpretations of the phenomena under investigation (e.g., *which* theory of consumer motivation; *which* etiology of skin cancer; *which* explanation of the nature of mass illiteracy). In basic research, the value of any study (or studies) can only be judged from a historical perspective.

The differential value of applied and basic research does not lie in any major differences in rigor of research methodology or validity of results (Bickman & Rog, 1998). Rather, from our perspective, the essential difference lies in the relative probability that results of research programs will contribute to the development of a broadly based explanatory theory or to a more limited exploration of some causal relation. To the extent that applied research is restricted to the examination of variations in a particular A-B relation, it is unlikely to uncover an explanatory principle that accounts for C-D and E-F, *along with* A-B. The applied social researcher is likely to limit the research explorations, for example, to sources of tension between Black and White people in contemporary American society. A basic research program, on the other hand, would more likely be concerned with the general phenomenon of ethnocentrism in intergroup relations, thus involving the investigation of Black-White relations, *as well as* inter-ethnic relations, international relations, and many other manifestations of the phenomenon (probably including interactions among lower animals within an evolutionary framework).

Many of those who are committed to basic social research contend that its long-run benefits, in terms of the alleviation of social problems, will be greater than those of applied research. However, just as the distinction between long term and short term is a relative one, the difference between applied and basic research contributions is a matter of degree. To the extent

that applied researchers are open to creative variations in their research problems, it becomes probable that they will serendipitously arrive at findings that will have broad theoretical impact. On the other hand, the more the basic researcher becomes involved in the permutations or combinations of a particular A-B relation, or the more committed he or she becomes to a minor theoretical point (that is, the closer one comes to the stereotypic version of the “ivory tower” scholar), the less likely it is that the research will contribute to a meaningful expansion of the explanatory power of the discipline, no matter how *inapplicable* the results of the research may be!

Basic Issues of Internal Validity

As we have indicated, the purpose of experimental design is oriented toward establishing causal relations and eliminating possible alternative explanations of research results (variations in scores on the dependent variable) that are unrelated to the effects of the treatment of interest (i.e., the independent variable). When an experiment is adequately designed, changes in the dependent variable can be attributed to variations in the treatment, which is systematically controlled and manipulated by the investigator. That is, response differences on the dependent variable can be accounted for by differences in the degree of exposure (or lack of exposure) to the experimental treatment. As will be discussed in Chapter 5, these differences may occur between measures on the same persons taken before and after exposure (*the pretest-posttest design*) or between measures on different groups that have been exposed to different conditions (*the comparison group design*). In either case, if obtained outcome differences can be attributed directly to and caused by the independent variable, the study is said to have internal validity. If factors other than the experimental treatment could plausibly account for the obtained differences, then the internal validity of the study is uncertain.

A **confound** is a type of extraneous variable that potentially affects the dependent variable in a way that cannot be separated from the effect of the independent variable. A research study is said to be confounded when there is reason to believe that obtained differences in the dependent variable would have occurred *even if exposure to the independent variable had not been manipulated*. For example, suppose we were testing the effectiveness of two different ads for a sports car. One of the ads is shown during an NFL football game and the other is broadcast during a daytime soap opera. A significant difference emerges, showing that the first ad is associated with more positive attitudes toward the sports car. However, there are large gender differences between the viewers of the two ads. Most of the viewers of the first ad were men, while women were overrepresented among the viewers of the second ad. So, the question is, did the ads cause the differences in liking for the sports car, or do men like sports cars more than women, regardless of advertisements? The presence of confounds such as this reduces internal validity and our ability to make causal inferences based on the results of such a study.

Potential sources of such invalidity that could affect almost any research program were identified and discussed by Campbell and Stanley (1963). The eight major **threats to internal validity** they discussed may be summarized as follows:

- 1 **History.** Differences in outcomes are attributed to events (unrelated to the experimental treatment) that occurred during the passage of time between the independent and dependent variables.
- 2 **Maturation.** Changes in outcomes are caused by changes in the internal or physical characteristics of the participants (e.g., growing older, becoming more tired, less interested). These are termed maturation effects, even though some representatives of this class (e.g., growing tired) are not typically thought of as being related to physical maturation.

- 3 **Testing.** Scores obtained in a second administration of the dependent variable measure are caused by participants having been exposed to the same measure previously.
- 4 **Instrumentation.** Differences in scores are caused by changes in the properties of the measurement instrument, rather than by changes in the participants being measured. Examples include the instrument being damaged or the measurement device failing.
- 5 **Statistical regression** (toward the mean). Outcome scores are subject to misinterpretation if participants are selected on the basis of extreme scores at their initial measurement session. Owing to unreliability of measures, at the later round scores will tend toward the mean score of the participants. (This threat to internal validity will be discussed more fully in Chapter 9.)
- 6 **Selection.** If outcomes for two or more groups of participants are being compared, differences are caused by selection procedures employed when participants were assigned to the groups.
- 7 **Mortality or attrition.** Selection procedures, treatment differences, or issues that yield different proportions of participants dropping out of the study may cause the observed differences between the groups in the final measurement.
- 8 **Selection-based interactions.** If participants were differentially (non-randomly) selected to serve in comparison groups, these specially selected groups may experience differences in history, maturation, testing, etc., which may produce differences in the final measurement.

An example of a simple research study may help to clarify how these threats to internal validity make it impossible to determine whether the independent variable is responsible for producing changes in the dependent measure. In this hypothetical study, the attitudes of a group of college students toward their school's administrators are measured. One week after the initial measurement, the same group of students is exposed to an administrator's speech (the experimental treatment), which advocates stricter discipline for students. The student group is then measured again on the attitude test. In this simple pretest-posttest design involving only one group, the researcher is interested in whether the speech communication changed attitudes expressed on the test. However, with this design, attitude changes (or lack of change) in the group occurring between pretest and posttest cannot be directly and unambiguously attributed to the effects of the communication because of the many uncontrolled extraneous variables that could provide rival alternative explanations.

For instance, during the passage of time between pretest and posttest, some event might have occurred on campus (*history*) that altered attitudes instead of the communication. The school authorities, for example, might have voted for a major tuition reduction. It should be noted here that although individual histories may cause individual changes in test scores, only *commonly experienced events*, which affect most of the group in the same way, will produce *systematic* changes between pretest and posttest scores. That is, individuals may have different (historical) experiences, some of which tend to increase attitude scores and some of which tend to decrease scores; however, across all individuals, the effects of these different experiences will cancel out, unless most of the individuals have the *same* experience, which exerts a consistent effect in one direction on all of their scores, thereby producing a discernible overall change in the total group's mean test score.

Similarly, the class of variables categorized as *maturation* effects could account for changes in attitude scores. It is possible, for instance, that as students become older and more experienced, they generally become more tolerant or accepting of authority figures. The longer the time period between pretest and posttest, the more likely it is that maturation or intervening historical events will provide plausible rival explanations for obtained attitude change.

A third rival explanation or threat to internal validity concerns *testing* and could arise from the possibility that pretesting produces sensitivity to the issue, which causes changes in responses to the second test. For some kinds of dependent variables, particularly achievement tests, test-taking practice effects usually lead to improved performance on a second measure, even if different questions are used. Unlike history or maturation, these *pretest sensitivity* or *practice effects* are more likely to provide plausible alternative explanations of results with *shorter* intervals between pretest and posttest.

Another type of internal validity threat, *instrumentation*, could occur if there were possible differences in scoring procedures employed between the pretest and the posttest. This would be especially likely if the attitude test was in the form of an essay exam (where scorers could shift standards between tests) rather than an objectively scored (e.g., multiple-choice) test. Whenever measurements are in any way dependent on subjective judgments of scorers or observers, as is often the case with many social variables, biases of the measure may produce instrumentation effects. For instance, in the above study, the researcher's assistants may have presented the pro-authority speech. If they were strongly personally committed to influencing the participants, and if they were scoring an essay-type attitude test under these circumstances, they would be more likely to detect pro-authority statements in the essays after the speech than during the pretest, thereby introducing scoring differences in favor of their own biases. Even with well-trained objective scorers, subtle differences in attitudes or conscientiousness over time may produce systematic instrumentation effects. Such effects can be controlled by requiring several unbiased scorers to agree in their judgments or by scoring pretests and posttests at the same time without letting the scorers know which is which.

Unfortunately, other rival factors cannot be so readily eliminated within the simple pretest-posttest design. Ideally, the factors of history, maturation, and testing could be controlled if we could measure each participant at the same time both with and without the experimental treatment. Then any differences between the two measures could be interpreted as owing to the effect of the treatment, and nothing else. As long as we are bound by the restrictions of Aristotelian logic, however, the best we can do is to measure two *different* groups of participants that are *equivalent* in terms of the effects of the rival factors. If one of these groups is then exposed to the experimental treatment and the other is not, then any changes in the "experimental" (or treatment) group above and beyond those that occur in the "control" group (the one not receiving the treatment) can be attributed to the experimental variable. Even though history, maturation, etc., may produce changes *within* the two groups, any difference *between* the groups can be accounted for by the experimental treatment that was administered to one group but not to the other.

Random Assignment and Experimental Control

As noted, the design of experiments is intended to eliminate threats to internal validity. The prototype of a good experimental design is one in which groups of people who are initially equivalent (at the pretest phase) are randomly assigned to receive the experimental treatment or a control condition and then assessed again after this differential experience (posttest phase). A graphic depiction of this *pretest-posttest-control group design* is presented in Figure 2.2.

The success of the two-group experimental design depends primarily on the assumption that experimental and control groups are equivalent on all factors except exposure to *different levels* of the independent variable. The ideal of comparability between two groups with respect to all variables other than the ones under experimental investigation may be met in either of two ways. Some variables may literally be "held constant," that is, be maintained at the same

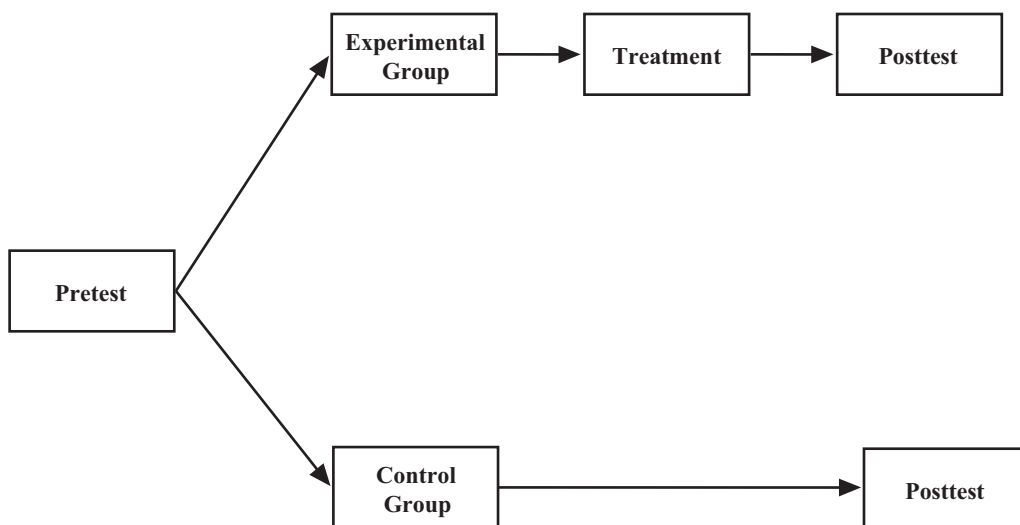


FIGURE 2.2 Diagram of a pretest-posttest control group design.

level for all participants in all groups (e.g., testing everyone at the same time of day, using the same experimenter for all participants, etc.). Other *unmeasured* variables may be allowed to vary randomly with the assumption that there will be no systematic differences between the sets of participants in the two comparison groups on these extraneous sources of variation. This is accomplished through the technique of random assignment of participants to groups.

Random assignment requires that all persons available for a particular research study be able to participate in either the experimental or the control group, and that only chance determines the group to which any individual is assigned. A coin (if there are only two groups) or random number generator may be used to assist in randomly assigning participants to groups. The faith in statistical chance to place participants into groups equally is bolstered as the number of participants in a study increases. The “law of large numbers” is crucial if randomization is to accomplish its goal. Without large numbers of participants available for assignment to conditions, randomization cannot deliver on its promise to produce equivalent groups. When enough participants are available, chance assignment assures that there will be no preexisting systematic differences linked to the independent variable (e.g., average age, birth order, sex, educational background, intelligence) between the groups at the onset of the study, and also that there is no reason to believe that they will experience any systematic differences during the research period, other than the treatment, whose delivery is controlled by the experimenter.

Without random assignment, myriad preexisting group differences would obscure the relation of the independent variable to the dependent variable relation under investigation. Random assignment represents the strictest form of experimenter control that can be imposed, as doing so rules out many, if not all threats to internal validity. For this reason, random assignment is required for any design to be considered a *true experiment*.

If any basis other than chance is used to assign participants to groups, then another threat to internal validity, *selection*, may account for differences on the dependent variable. Particularly problematic is any selection procedure that permits participants themselves to determine to which treatment they will be exposed. Such *self-selection* on the part of the research participants makes it impossible to tell whether the treatment affected the dependent variable scores or whether differences in scores were determined by the personal characteristics of the individuals who chose to expose themselves to, or avoid, the experimental treatment. For example, in educational research

dealing with remedial teaching programs, it is unwise to expose volunteers to the program and then to compare their progress with that of a “control” group of students who declined to take part in the program. If any differences in achievement do show up, it is not possible to determine whether they were brought about by the remedial program or whether characteristics of the volunteer participants, such as greater motivation to improve, were responsible for performance differences. Only volunteer participants who were randomly assigned to a no-remedial-treatment condition could provide an equivalent control group for comparison with the experimental group.

The self-selection of participants to experimental or control conditions can cloud the interpretability of a study, but other forms of selection artifacts also can arise when randomization is not employed, and these sometimes can prove even more difficult to recognize than simple self-selection. Unfortunately, in some research circumstances, non-random selection is nearly impossible to avoid. Under such circumstances, caution must be exercised in interpreting research findings. Let us expand our example of an experimental testing program to illustrate this point.

Suppose that the government developed a program designed to offset the academic deficiencies of children whose parents fell below federal poverty guidelines. Only poor children, in other words, were eligible for this program. How would the researcher determine whether the educational program was succeeding? A pretest-posttest-control group experimental design would seem ideal for the investigators’ purposes. But who would constitute the control group? From our earlier discussion, it is clear that one would not compare the academic achievement of treatment group participants with that of children who were eligible for the program, but who had failed to volunteer for it. Too many differences already exist between volunteers and nonvolunteers to trust such a comparison (Rosenthal & Rosnow, 1975).

Alternatively, one could sample from a group of children who were not eligible for the program. However, these “control” group children likely would be from a higher economic stratum than that of the experimental group, and we already know from years of educational research that there is a positive relation between socioeconomic status and academic achievement. Thus, a comparison of this type would almost surely show large initial (pretest) differences between the groups, in favor of the (wealthier) control group participants. Such a preexisting difference would reflect the non-random selection problem operating here.

Another possibility would be to systematically search for children in the control group whose pretest scores exactly matched those of children in the experimental condition. If a suitable one-to-one match could be found for every participant in the experimental group, the average pretest scores of the two groups would be identical. Whereas this might seem an appealing solution to some, it is fraught with difficulties because it implicitly assumes that the histories, maturation rates, etc., of the two samples will be identical over the course of the study. This is a risky assumption at best, and in light of considerable research (Cronbach & Snow, 1977; Trochim, 1984, 1986), palpably unlikely in the present circumstances. We know that children of higher socioeconomic status are academically advantaged relative to their poorer peers, and that this “achievement gap” widens as time goes by. This *selection × maturation interaction* threat to internal validity often is difficult to recognize, and even more difficult to offset. Likewise, if the histories of the two groups differed systematically over the course of the experiment (a likely possibility if the treatment extended for any period), we will encounter an interaction of *selection × history* – effects that can bias our ability to infer causation just as surely as would a simple selection bias.

Solutions to the problems introduced by selection artifacts are not simple. Generally, complex statistical techniques are employed in attempts to assess, if not offset, the biases that can occur when the assignment of participants to conditions is not random (see Chapter 9 for a more complete discussion of quasi-experiments). It should be kept in mind that these statistical “fixes” are not ideal, but rather akin to a salvage operation. Clearly, if random assignment can be accomplished, it should be.

Participant Loss

The advantages of random assignment for internal validity are assured if the initial equivalence of the experimental and control groups can be maintained throughout the study. Unfortunately, research involving human participants is never free of the possibility that at some time between random assignment to groups and the posttest, some may drop out of the experiment and become unavailable for final testing. If the groups are initially equivalent, and if this dropout rate (referred to as the more morbid as participant mortality) occurs randomly across groups, then the loss of participants should not be any different for the experimental and control groups, and so does not pose a serious threat to internal validity. If we have pretested all participants at the beginning of the experiment, we can check the pretest scores for those participants who have dropped out. If those scores are essentially the same for dropouts from both the experimental and the control groups, we may have some confidence that the two groups are still equivalent except for effects of the experimental treatment.⁴

If there is something about the experimental treatment that enhances or otherwise affects the chances of participants dropping out of the study, a serious problem is introduced because the initially equivalent groups may become differentially selected groups by the time of final testing. For instance, if after random assignment to research groups, participants in the experimental treatment group learn that they are going to be faced with a task that involves a lot of hard work or effort as part of their participation, many of them may find excuses for discontinuing their involvement in the experiment. Thus, at the time of post-treatment testing, the experimental group would be composed largely of unusually hard-working or unusually compliant participants, unlike the control group that had not been exposed to the same selective pressures.

Researchers must be aware that participant selection is a potential problem not only at the initial stage of assigning participants to groups but throughout the study, and should attempt to design the experiment so that differences between the treatments received by experimental and control groups will not introduce differential tendencies to drop out of the experiment. In the example given here, for instance, the experimenter could see to it that both experimental and control groups were led to expect that they might all have to work hard (even though only the experimental group would actually do so). This procedure would not prevent participants from dropping out of the study, but it would help assure that such dropouts were equally distributed between experimental and control conditions.

Sometimes participants are dropped from an experiment by the researcher, either during the course of the experiment or at the time of data analysis. For instance, upon identifying one or two participants who show unusual performance after the experimental treatment, the researcher might examine the background information for these participants, conclude that they were highly atypical, and thereby justify the removal of their data from the final analysis. If the participants in the control group are not also examined for similar atypical cases, the researcher has applied a selection rule that may result in an apparent “treatment effect” that would not have appeared if all the data had been included.

Reprise

To review the eight threats to internal validity, consider examples using a common scenario, as presented in Table 2.1. Setting up this study, suppose that using a pretest-posttest control group design, a therapist creates a treatment plan to help people who are habitually unable to enjoy the antics and delights of clowns, and have been diagnosed with clown phobia. [This clinical problem is called coulrophobia (Flora, 2006).] To measure the level of fear and anxiety toward

TABLE 2.1 Threats to internal validity as illustrated with a clown phobia study

<i>Threat to internal validity</i>	<i>Example</i>
1 History	While all experimental group participants are sitting in the waiting room for the intervention to begin, the television in that room airs breaking news that a disgruntled clown was just convicted of arson in burning down a circus tent. As the control group did not view the news, higher scores are found on the posttest for the experimental group.
2 Maturation	The control group is administered the clown phobia posttest immediately, but the experimental group receives it one year after the therapy. As participants in the experimental group became older and wiser, and thus outgrew their fear of clowns, lower posttest scores were reported than control participants.
3 Testing	Participants in the experimental group were more likely to have answered a clown phobia questionnaire as part of a previous study by another researcher. Encountering the same questions again, experimental group participants rush through the questionnaire very quickly without paying attention.
4 Instrumentation	For the control condition, the researcher administers the full 100-item posttest. For the experimental condition, the researcher fails to make photocopies of the last page of the questionnaire, and accidentally administers an 80-item version. Consequently, the total posttest score in the experimental group appears to be lower than in the control group.
5 Statistical regression	As the researcher is interested in only the most severe clinical cases, only participants with the highest clown phobia scores on the pretest are selected to receive the therapy. Thus, an overall pretest to posttest reduction in clown phobia is observed among treated participants.
6 Selection	Participants are allowed to choose which group they wish to participate in, with those most motivated to reduce their fear of clowns tending to select the experimental group. Results show that the experimental, in comparison to the control, group showed lower posttest scores.
7 Mortality	As part of the exposure treatment requires putting on aesthetically questionable clown make-up, participants in the experimental group are more likely than those in the control group to drop out before completing the posttest.
8 Selection-history interactions	Allowed to select the group, more motivated participants request to serve in the experimental group. Also because of this motivation, they are more likely to watch, in their spare time prior to returning for the posttest, online videos of friendly clowns making animal balloons.

clowns, both the experimental and control groups complete a clown phobia questionnaire at pretest and at posttest. Between both test administrations, only the experimental group receives the treatment of exposure therapy involving desensitization to clown paraphernalia and activities, which is designed to reduce their unrealistic terror of these circus professionals. Table 2.1 shows how the various threats to internal validity could operate in this research context.

Other Sources of Invalidity

Complete control over experimental treatments is necessitated by the fact that inferences drawn from studies using a control group are acceptable when the treatments to which the experimental

and control groups are exposed differ *only* on the variable under consideration (i.e., all other conditions are “held constant”). For example, in physiological research dealing with the effects of surgical ablation of the parts of the brain, some animals are subjected to the full surgery (the experimental group) while others go through a “sham operation,” in which they undergo all the phases of the surgery *except* the actual removal of the brain section. The sham operation control is used to assure that differences between experimental and control animals are attributable to effects of the surgical ablation and not to experiences associated with general operative procedures, such as anesthesia or postoperative shock. Similarly, a researcher interested in the effects of anxiety-producing situations must be sure that the conditions to which the experimental and control groups are exposed differ only in induced anxiousness and not in such extraneous variables as the relationship between participant and experimenter.

The fallible observer. One source of internal invalidity that can threaten an experiment even when random assignment has been used is derived from the researchers’ *expectations* about how the experiment will (or should) turn out. In Chapter 7, we will discuss research indicating how such expectations can influence in relatively complex and subtle ways the nature of the experimenter-participant interaction. But more direct experimenter effects may influence the outcome of a study independent of any effects upon participant responses. Such effects are generated by *unintentional* differences in criteria applied by observers to participants in different groups, *unintentional* misinterpretation of participants’ responses, *unintentional* errors of data recording and analysis, and the like.⁵ (Such factors would be considered “instrumentation” effects in our earlier discussion of threats to internal validity.)

A valuable example of the effects of observer bias on the (mis)recording of participants’ responses was provided by Kennedy and Uphoff (1939). These experimenters asked individuals who were classified based on their belief or lack of belief in extrasensory perception (ESP) to take part in a test of ESP ability. Using a standard card-guessing task, the participant-observers were to “transmit” one of four symbols portrayed on each of a set of cards to another individual and to record the guesses made by this “receiver.” The persons whose ESP was being tested were, in fact, experimental accomplices, and the “guesses” they made were prerecorded. The principal dependent measure of this investigation was the number of times the participant-observers erred in recording the accomplice’s responses. Of 11,125 guesses, only 126 (1.1%) were misrecorded. Given the magnitude of the recording task, an *error rate* of 1.1% would not seem overly bothersome. However, closer examination of the *direction* of these errors as a function of the individual student-observer’s beliefs in ESP proved revealing. Of the 57 errors tending to increase telepathy scores, 36 (63%) were committed by believers in ESP. Conversely, 18 of the 27 errors (67%) lowering telepathy scores were made by the nonbelievers. These errors, then, were influenced by observers’ prior beliefs. They clearly were not random, though they were admittedly small, as is typical of findings in this area of research.

Controlling variables. The ideal of *ceteris paribus*, that there be no differences in conditions to which comparison groups are exposed except the experimental variable under consideration, is seldom perfectly met in practice. The experimental and control groups can never be treated precisely alike in every detail; some minor variation can be ignored as irrelevant. However, it pays to be aware of details – today’s irrelevancy may become tomorrow’s major breakthrough. At one time, for example, researchers paid little attention to the correspondence between a persuasive source’s group membership and that of the group being influenced. Now, group identity is recognized as an important factor in a source’s power to influence others (Turner, 1991).

The need for careful control over experimental variables has driven some social researchers to limit themselves to laboratory research and ritualized research techniques. However, it is our hope that greater awareness of the principles that underlie research methods will encourage

investigators to apply their ingenuity to achieving controlled variations in a variety of potential research settings.

Basic Issues of External Validity

The design issues we have considered thus far in the chapter have been concerned almost exclusively with the internal validity of a research study. In many ways, internal validity is the *sine qua non* of good experimental research. The essence of experimental design is to control the assignment of participants to treatment groups and the conditions of treatment delivery in such a way as to rule out or minimize threats to the internal validity of the study. In this way, any differences observed in the dependent measures can be traced directly to the variations in independent variables introduced by the experimenter. For experimental as well as nonexperimental strategies, there may arise questions about the validity of interpretations of relations obtained in any given study, particularly their applicability or generalizability beyond the results of the specific study. These concerns constitute issues of external validity, which can be further divided into questions of (a) generalizability of operationalizations and (b) generalizability of results to other places and participant populations (Cook and Campbell, 1979; Shadish et al., 2002).

Generalizability of Operationalizations

Concerns over the validity of operations refer to the correct identification of the nature of the predictor and outcome variables and the underlying relation between them, that is, the extent to which the operations and measures embodied in the procedures of a particular study reflect the theoretical concepts that gave rise to the study in the first place. Threats to this form of validity arise from errors of measurement, misspecification of research operations, and, in general, the complexity of stimulus features that constitute our variables.

The complex constructs characteristic of social research are marked by what Aronson et al. (1998) called “multiple meaning,” that is, their impact on participants may be due to any of several factors inherent in the research situation, many of which may be completely irrelevant to the conceptual variables of interest. In such cases, researchers cannot be sure that any effects obtained reflect the influence of the one construct they were attempting to represent in their experimental operations. Confidence in interpretation is enhanced if a series of experiments is conducted in which the conceptual variable is represented by several *different* operations that vary as much as possible, having in common only the one basic factor of (theoretical) interest. When different techniques produce the same result, attributing the effect to the common conceptual variable is substantiated.

As an illustration of this process of identifying the “active ingredient” of our manipulation, consider a classic experiment undertaken by Aronson and Mills (1959) to test the hypothesis that undergoing a severe initiation to join a social group enhances the attractiveness of that group to the research participants. In their original experiment, Aronson and Mills convinced female college students that they needed to pass an “embarrassment test” to qualify to participate in a series of group discussions on various sensitive topics. In one version of the experimental procedures, the initiation test was relatively mild; but the severe initiation condition required that the participant read aloud (in front of a male experimenter) a considerable amount of sexually explicit material and a list of obscene words. As predicted, those participants who underwent the severe initiation were more favorable in their later evaluations of the group discussion (which was purposely designed to be quite dull) than were the participants in the mild initiation condition.

Aronson and Mills (1959) designed their study as an operational representation of “effort justification,” a concept derived from dissonance theory. The basic idea being tested was that the participant needed to justify effort expended in qualifying for the group discussion, and they could do so by evaluating the goal of their effort – the group discussion – as worth that effort. The greater the effort involved, the greater the need for justification and hence, the greater the need to value the group experience positively. When the experiment was published, however, several alternative explanations for the effect of initiation severity on attraction ratings were proposed. Most of these alternatives revolved around the fact that the Aronson and Mills’ “effort” manipulation procedures involved a potentially sexually arousing experience. If that were the case, then the carry-over of sexual excitement to evaluations of the group discussion, rather than effort justification, may have accounted for the positive effects of the severe initiation condition (Bryant & Miron, 2003).

To rule out this alternative interpretation, Gerard and Mathewson (1966) designed a replication experiment with a different operationalization of effort expenditure. In their experiment, electric shock, rather than reading obscene materials, was used as the severe initiation experience, and the shocks were represented as a test of “emotionality” rather than as an “embarrassment test.” Having thus removed sexual arousal as a component of the experimental manipulations, Gerard and Mathewson obtained results consistent with Aronson and Mills’ (1959) original study: participants who underwent painful shocks to join a dull group evaluated that group more positively than did participants who underwent very mild shocks. This replication of the basic effect with different experimental operations confirms, at least indirectly, the effort-justification interpretation of the original study and provides support for the external validity of the initial experimental manipulations. The Gerard and Mathewson experiment represents a **conceptual replication** of the earlier Aronson and Mills (1959) study. The role of such replication studies in the validation of social science theory will be discussed in more detail in Chapters 8 and 20.

Generalizability of Results to Other Places and Populations

Once a research study has been completed, the investigator usually is interested in reaching conclusions that are generalizable across people and settings. Results that are applicable only to particular persons at a particular time or place are of little value to a scientific endeavor that aims at achieving general principles of human behavior. Generalizability refers to the robustness of a phenomenon – the extent to which a relation, once identified, can be expected to recur at other times, places, and populations under different environmental conditions. Threats to this form of external validity arise from possible *interaction* effects between the treatment variable of interest and the context in which it is delivered, or the type of participant population involved. A finding lacks external validity if the nature of the effect of the predictor variable would be reduced or altered if the setting or the participant population were changed. Because so many of the laboratory studies in social sciences are conducted with college students as participants, it has been suggested that the truth of the relations we observe may be limited to that particular population. If it happens that college students – with their youth, above-average intelligence, health, and socioeconomic background – respond differently to our variables than do other types of people, then the external (but not internal) validity of our findings would be suspect. But this hypothesis about population differences is subject to the same rules of evidence as any other hypothesis (Sears, 1986).

Just as the validity of measurement depends upon an adequate representation of variation in the theoretical concept, participant generalizability depends upon the extent to which the

participants included in a study represent the potential variability of humans. We have already discussed the principle of random assignment of participants to treatment conditions, but **random selection** requires that chance determines selection of participants for a study and is used to assure generalizability of results from the sample to the entire population of relevant persons. (In Chapter 12 we consider several different forms of respondent selection.) Avoid confusing random selection and random assignment, as the only similarity is that chance is used as an ally. The ground rule of random selection is that all persons *in the population of interest* are equally likely to be included in the research sample, whereas random assignment requires that persons *in the research sample* are equally like to participate in the different groups within a study. Another critical distinction is that the purpose of random selection is to increase a study's external validity, whereas random assignment is to increase a study's internal validity.

Random selection to yield a representative sample is a rarely realized ideal in social research. Limited resources and participant availability make it impossible, and in many cases, the necessity for keeping experimental and control groups equivalent further limits the participant pool. Under the typical volunteer participant conditions, the usual procedure is to randomly sample from some readily available population and then to generalize to those persons who cannot be assumed to be systematically different from those who took part in the research study. This still places severe limitations on generalizability of findings – research results obtained from college sophomores can hardly be automatically assumed to apply to a population that includes grade school dropouts and people over 60 – but the cost and impracticality of true random sampling usually make this limitation a necessity. Crowdsourcing approaches (e.g., Amazon's Mechanical Turk) are becoming popular ways of gathering samples of participants who are not necessarily college students, and who can represent a wide range of socio-demographic features. (We discuss these approaches in Chapter 8.)

One way to overcome the lack of generalizability of most research studies is to repeat essentially the same design with different populations of persons as they become available. To the extent that research results are reproduced with different types of persons, generalization across all persons becomes more convincing. Even a *failure* to replicate findings with different populations adds to our understanding of the phenomenon being investigated, by identifying the limitations of the effects of the variables under study.

Apart from generalizability of the sampled participants to the population of interest, the reproducibility of any research findings may be limited to the conditions under which the phenomenon was studied. *Condition replicability* involves two aspects: the internal conditions of the research participants, across time, and the external physical and social environment in which the research is carried out. Too often, replication of research studies is conducted with the aim of reproducing, in precise detail, the exact conditions of the original research. Apart from the inevitable futility of such an approach to replication, it is not consistent with the aim of identifying principles and relations that have some survival value across heterogeneous circumstances. Conclusions are of limited value if they cannot satisfy this criterion of robustness in the face of altered research conditions.

Conclusion

Enhancing the internal validity of research operations generally involves the use of highly controlled experiments, in which participants from a restricted group (e.g., introductory psychology students) are assigned randomly to study conditions. This tactic helps enhance internal validity but a most serious limitation to generalizability of research results. The laboratory setting represents a unique and isolated social experience, in which the researcher exercises

unusual control over the environmental conditions to which the participant is exposed, and which limit the participant's available choices of behavior. In many respects this is the strength of the scientific approach to understanding social phenomena, for until simple principles are isolated and identified, the complex gamut of social relations will remain unfathomable. However, the unusual relationship that occurs between researcher and participant, and the participant's awareness of being an object of scientific investigation, may produce phenomena that are unique to the laboratory setting. Until the implications of such research are examined under broader social conditions, any generalized conclusions drawn are at best tentative.

Notes

- 1 The word "program" is used loosely here to refer to accumulated evidence relevant to a particular theoretical position, whether planned and executed by a single researcher or research team, or derived from the uncoordinated efforts of independent researchers working in the same area.
- 2 Lachenmeyer (1970) makes a strong argument for the role of experimental methodology in all phases of theory construction, including the development of observational techniques and measurements, fact-finding exploratory research, and the verification of the existence of hypothesized relations. In our view, however, the unique features of the experiment make it most useful as a hypothesis-testing device.
- 3 In the social research literature, the use of statistical significance testing has come under challenge (Hunter, 1997), but for many reasons, it still represents current practice (Abelson, 1997; Estes, 1997).
- 4 The problem with this technique is that it relies on *accepting* the null hypothesis of no difference between the drop-outs and those who remained in the experimental groups. A more thorough approach would compare groups on a host of theory-relevant variables to ensure that the drop-outs are not in some ways systematically different from the remainders on important, theory-relevant variables.
- 5 The word *unintentional* is stressed here because such effects can occur even when the experimenter makes an honest effort to avoid them.

References

- Abelson, R. P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science*, 8, 12–15.
- Agnew, N. M., & Pike, S. W. (1969). *The science game. An introduction to research in the behavioral sciences*. Englewood Cliffs, NJ: Prentice-Hall.
- Aronson, E., & Mills, J. (1959). The effect of severity of initiation on liking for a group. *Journal of Abnormal and Social Psychology*, 59, 177–181.
- Aronson, E., Wilson, T., & Brewer, M. B. (1998). Experimentation in social psychology. In D. Gilbert, S. Fiske, & G. Lindsey (Eds.), *The handbook of social psychology* (4th ed., Vol. 1, pp. 99–142). Boston, MA: McGraw-Hill.
- Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173–1182.
- Bickman, L., & Rog, D. J. (Eds.) (1998). *Handbook of applied social research methods*. Thousand Oaks, CA: Sage.
- Brewer, M. B. (2000). Research design and issues of validity. In H. Reis & C. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 3–16). New York: Cambridge University Press.
- Bryant, J., & Miron, D. (2003). Excitation-transfer theory and three-factor theory of emotion. In J. Bryant, D. Roskos-Ewoldsen, & J. Cantor (Eds.), *Communication and emotion: Essays in honor of Dolf Zillmann* (pp. 31–59). Mahwah, NJ, US: Erlbaum.
- Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Chicago, IL: Rand-McNally. Reprinted as *Experimental and quasi-experimental designs for research*. Chicago: Rand-McNally, 1966.
- Cohen, J. (2003). A power primer. In A. E. Kazdin (Ed.), *Methodological issues & strategies in clinical research* (3rd ed., pp. 427–436). Washington, DC: American Psychological Association.

- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago, IL: Rand-McNally.
- Cronbach, L. J., & Snow, R. (1977). *Aptitude and instructional methods*. New York: Irvington.
- Estes, W. K. (1997). Significance testing in psychological research: Some persisting issues. *Psychological Science*, 8, 18–20.
- Flora, C. (2006). No laughing matter. *Psychology Today*, 39, 10–11.
- Gerard, H. B., & Mathewson, G. (1966). The effects of severity of initiation on liking for a group: A replication. *Journal of Experimental Social Psychology*, 2, 278–287.
- Hallahan, M., & Rosenthal, R. (1996). Statistical power: Concepts, procedures, and applications. *Behaviour Research and Therapy*, 34, 489–499.
- Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York: Guilford Press.
- Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science*, 8, 3–7.
- Kennedy, J. L., & Uphoff, H. F. (1939). Experiments on the nature of extrasensory perception: III. The recording error criticism of extra-chance scores. *Journal of Parapsychology*, 3, 226–245.
- Lachenmeyer, C. W. (1970). Experimentation -- a misunderstood methodology in psychology and social psychological research. *American Psychologist*, 25, 617–624.
- Mook, D. G. (1983). In defense of external invalidity. *American Psychologist*, 38, 379–387.
- Rosenthal, R., & Rosnow, R. W. (1975). *The volunteer subject*. New York: Wiley-Interscience.
- Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, 51, 515–530.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton, Mifflin and Company.
- Smith, E. R. (2000). Research design. In H. Reis & C. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 17–39). New York: Cambridge University Press.
- Trochim, W. (1984). *Research design for program evaluation: The regression-discontinuity approach*. Beverly Hills, CA: Sage.
- Trochim, W. (1986). *Advances in quasi-experimental design and analysis*. San Francisco, CA: Jossey-Bass.
- Turner, J. C. (1991). *Social influence*. Milton Keynes: Open University Press.

3

MEASUREMENT RELIABILITY

Measurement is a form of operationalization in which abstract concepts are *translated* into observable and quantifiable information or data. In social research, various measurement techniques are used to assess differences among the people or units being studied. The objects of measurement can be individuals reporting on themselves, observations of behaviors, characteristics of situations, or records of past events. We adopt a classic measurement theory orientation, also known as psychometric theory, which is focused on any observation, ability, state, trait, or behavior. The classic theory approach assumes a consistent, stable underlying state that reflects the construct we seek to quantify.

Although all translations (that is, measures) are imperfect, individual measures vary in the adequacy with which they characterize the underlying conceptual variable of interest. Some measures come closer than others to representing the true value of the concept, in part because they are less susceptible to sources of systematic error or random fluctuation. The quality of a given measure is expressed in terms of its reliability and validity. Briefly, **reliability** is the consistency (or replicability) of a measurement instrument. It says nothing about the meaning of the construct, or whether the construct is adequately reflected by the measure – that is, the role of **validity**, the degree of relationship, or the overlap, between a measurement instrument and the construct it is intended to assess. In this chapter we review methods for assessing the reliability of psychological measures; evaluation of measure validity will be addressed in Chapter 4.

Classical Measurement Theory

True Score and Error

The concept of reliability derives from classical measurement theory, which assumes that the score obtained on any single measurement instrument represents a combination of the *true score* of the construct being measured and random *errors* that lead to fluctuations in the measure, obtained from the same source at different occasions (Gulliksen, 1950). The standard classical test theory formula ($O = T + E$), modified slightly to fit our conceptualization of reliability, is better expressed as follows:

$$O = T + \Sigma E_{r+s}$$

where O = **observed score** or score obtained on a measurement instrument (e.g., math test, behavioral checklist, or attitude scale),

T = true score,

and ΣE_{T+s} = the combination of random and systematic errors that merge with true score to produce the observed score.

The standard formula usually lists only random error; it does not consider systematic error, or bias as it is often called. The bias error components are pooled with random error in the formula for observed scores. However, combining random and systematic error is misleading. Although both forms of error affect observed scores, they do so in different ways. We have modified the original formula to indicate that both types of error affect observed scores.

The **true score** is the replicable feature of the concept being measured. It is not “true” in the sense that it is a perfect or necessarily valid representation of the underlying construct. “True” in the classical test theory sense signifies replicability, nothing more. It is the component of the observed score that would recur across different measurement occasions in the absence of error. The true score represents the reliable portion of the observed measurement across infinite potential measurements. True score has to do with measurement consistency, the measure’s repeatability across many observations (Raykov & Marcoulides, 2011). On any single measurement occasion, we know only a participant’s observed score, but the formula contains two unknown elements – the true score and error – making this equation unsolvable. To estimate the error component, we must have replication of measurement.

A simple example may help clarify the meaning of true score. Often, when using a standard bathroom scale, we will notice our observed weight fluctuating, sometimes even when weighing ourselves twice almost simultaneously. Has our weight changed over the intervening seconds? Probably not, but factors extraneous to “true” weight have varied from the first to the second weighing (the tension of the scale’s spring, the placement of our feet on the scale, etc.). These extraneous factors are error. They really do not have anything to do with how much we weigh. They degrade the relationship between true and observed scores. The greater the proportion of error, the less the observed score reflects the underlying true score, and the more unreliable the measure is said to be – because it is more greatly influenced by error.

To understand the relationship of reliability and measurement error, it is important to distinguish between random and systematic error. **Random error** is attributed to unexpected events that artificially influence observed scores in a *nonsystematic* way. Examples include inadvertent misrecording (a tired coder may misrecord values – sometimes they are too high, at other times too low), a faulty spring on a bathroom scale sometimes overestimates and other times underestimates people’s weight; the mood induced by communicating with one’s parent earlier that day may result in a more positive mood for some respondents, whereas others might have a more negative mood than usual. Chance events, by definition, are nondirectional on average, that is, across many replications, negative and positive errors theoretically cancel each other. According to classical test theory, if the sample size is large enough, random errors across respondents will sum to zero.

This is not to suggest that random error is harmless, however, because such errors not only reduce the accuracy of measurement, they also affect the measure’s sensitivity to detecting differences in observed score between different groups of participants. Random error increases variability or spread of scores from the mean, and increased variability reduces the power of tests of statistical significance. Obscured by random error, larger mean differences between groups are needed before the true differences between groups can be judged real or trustworthy (or statistically significant, above random chance).

In contrast to random error, **systematic error** is attributed to unanticipated and irrelevant factors that tend to inflate or deflate observed scores *systematically*. Examples of systematic error include a researcher who smiles and is overly friendly toward participants as they are entering the research laboratory (upwardly biasing their self-esteem), forgetting to ask participants to remove their shoes prior to getting on a scale (weight is overestimated across all participants), or a heavy snow storm on the day of the study (consistently inducing more negative mood states than usual across respondents). Systematic error shifts most, if not all, scores in the same direction. Because systematic error (bias) is not random, it does not cancel between groups; rather, it either exacerbates or mitigates differences that exist. The outcome of biased measurement depends on the direction of the bias. If the difference between groups becomes smaller because of systematic error (bias), it is less likely to be statistically significant than if the measure were not biased. If the difference between groups becomes larger because of biased measurement, it is more likely to be statistically significant. Biased measures are prone either to Type I or Type II errors, either of which should be avoided by developing or adopting measures of high reliability.

An example of a hypothetical experiment will help to clarify the conceptual distinction between random and systematic error. Suppose an experimental treatment (e.g., positive compliments) is designed to increase the likelihood that people will succumb to a persuasive communication. In the experimental group, participants are praised privately by a confederate that their performance on a task indicated they were highly intelligent. The control group participants are given no feedback on their performance. Both groups then read a persuasive message that argues for future enrollment increases at their university. The dependent measure is an attitude scale that assesses their extent of agreement with this message. Scores can range from 0 to 20, with higher scores indicating greater agreement.

The true scores of the treatment (praised) and control groups are presented in the first two columns of Table 3.1. (The values in these columns depict idealized results from the hypothetical scenario; true scores and the sources and degree of error are unknowns.) A statistical test of the difference between mean true scores of these two groups discloses a significant difference, with the treated participants (Column 2) demonstrating greater agreement with the message than the control participants (Column 1).

TABLE 3.1 Results of a hypothetical experiment with true scores and error components

Participant #	(1) Control group true score	(2) Treatment group true score	(3) Control group (random error)	(4) Treatment group (random error)	(5) Control group true score + (random error)	(6) Treatment group true score + (random error)	(7) Treatment group true score – (systematic error)	(8) Treatment group true score + (systematic error)
1	8	12	2	6	10	18	6	18
2	6	10	–1	–6	5	4	4	16
3	7	9	4	–1	11	8	8	10
4	9	10	–4	4	5	14	6	14
5	9	11	8	2	17	13	9	13
6	12	14	–6	–7	6	7	7	21
7	6	12	4	1	10	13	11	13
8	10	13	–8	5	2	18	8	18
9	4	12	–2	1	2	13	11	13
10	3	10	3	–5	6	5	5	15
Sum	74	113	0	0	74	113	75	151

Suppose, however, that our attitude-measuring device contained random error, which either raised or lowered actual test scores for each participant. Overall, however, being random (or unbiased), the error averaged out to 0 for each group. Then the observed scores that we obtained would be those derived through an addition of Columns 1 and 3 (true score + random error) for the control group, and Columns 2 and 4 for the treatment group. Notice that the same degree of “error” was added to each group’s scores (the column sums are both 0). Thus, the respective column sums remain the same. Yet, upon reanalysis, the same statistical test that had disclosed a significant difference between Columns 1 and 2 now disclose no significant difference between the groups (i.e., when comparing Columns 5 and 6). This reversal of fortune is attributable to the increased within-group variability of scores when the random error components are added to each column. With these results, we would be forced to conclude, incorrectly, that our manipulation had no effect on persuasion. This form of erroneous conclusion is termed a Type II error (Chapter 2).

To illustrate systematic error (bias), suppose now that all participants in the experimental condition were given the attitude test in an uncomfortably hot room, putting them in a negative mood. This mood influenced their responses in such a way that the error scores of the treatment group were all negative, i.e., systematic error *decreased* all the scores of this group only. (Note that all error scores in Column 4 are negative.) The ratings of the control group participants remained unaffected. This systematic deflation of the observed treatment group’s scores (Column 7) in the direction of the observed control group scores (Column 2) reduces the difference between groups, even though the true scores were significantly higher for the experimental group. The results would lead the researcher to conclude that the “hot room” treatment was unsuccessful (another instance of a Type II error).

Now, suppose instead that the participants were given a chocolate sundae to eat while completing the measure (and assume everyone loves chocolate sundaes). In this case, this testing circumstance of the experimental group might bias the results in a *positive* direction, confirming the study’s hypothesis. To show how this would change the conclusions drawn from the study, the systematic error score (Column 4) is now added to each treatment participant’s true score (Column 2), but the control group participants are unaffected by the bias. Results now reveal a larger difference in scores between the control group (Column 1) and the treatment group (Column 8), yielding a *more* statistically significant finding (from $p < .05$ to $p < .001$). Had the original difference in true scores between the control (Column 1) and treatment (Column 2) groups not been statistically significant in the first place, this direction of systematic error would have increased the likelihood that results would have achieved statistical significance. Recall (Chapter 2) that erroneously finding a statistically significant result that is not truly attributable to the treatment is termed a Type I error.

Assessing Reliability

Systematic errors in measurement become part of a respondent’s “true score” on that observed measure, and hence affect its validity as a measure of the conceptual variable of interest. As we have illustrated, systematic error (bias) can result in either Type I or Type II errors. Depending on the direction of the bias, it fosters conclusions of either a stronger or a weaker difference between groups. Random errors, on the other hand, affect the measure’s reliability. Random error lessens the chances of finding a true difference between groups when a true difference exists. As such, random error fosters Type II errors. When measures are taken on a large number of individuals with a given instrument, the variability in obtained scores is due partly to differences among those individuals in their true scores on the measure, and partly to random and systematic fluctuations. Technically, the reliability of a measure is defined as the proportion of the total

variance in observed scores that is due to true score variability. A perfectly reliable instrument would be one in which this proportion is equal to 1.00, or in which true score equaled observed score ($\text{Observed} = \text{True} + 0.00$). A perfectly unreliable score, on the other hand, would be one in which the observed score equaled the sum of the error components, and true score contributed nothing to the observed score ($\text{Observed} = 0.00 + \text{Error}$). It is hard to imagine a measure this bad, but it could exist theoretically. The important take-home message of the formula for classical test theory is that instruments are fallible in practice because they are susceptible to incidental and irrelevant sources of random and systematic error.

Although the technical definition of reliability presented in the prior paragraph is standard, the actual meaning of the term reliability varies, depending upon how it is assessed. Reliability can refer to the degree to which respondents' scores on a given administration of a measure resemble their scores on the same instrument administered at some later point in time – or the extent to which two judges observing the same behavior produce the same ratings of the behavior. If the test-retest scores tend to be very similar (i.e., highly interrelated), the measure is said to be reliable. Or, if parallel forms of a test – two forms of the test that are thought to measure the same construct – are highly correlated, the test is said to be reliable. However, reliability has also come to mean the degree to which the set of items or questions *within* a particular multiple-item scale are interrelated. These two types of reliability are termed *internal consistency* and *temporal stability*. These features of reliability are important and should be considered when evaluating the quality of any measurement instrument.

Internal consistency. The question of internal consistency is concerned with the extent to which the components (e.g., individual items, observations) of a measuring instrument are interrelated. The idea of internal consistency is usually applied to measures that consist of a set of items, as often are found in ability tests or attitude scales. It is assumed that all the items of the scale measure the same underlying construct. The same logic is applied when the “measuring instruments” are human observers, or judges. In this case, the question is, “Have the judges seen the same thing, as inferred from their giving more or less identical scores to the same events or behaviors?” This third type of interrater reliability is covered in Chapter 12. The answer to the question is assessed by the extent to which the observers' observations overlap, or correlate. If observers assessing the same phenomenon (say, a group discussion) use different criteria, then combining their individual observations into an overall summary score is logically indefensible. Similarly, if the items that purportedly constitute a scale assess a variety of different constructs (i.e., if the scale is *multidimensional*), then there also is little to justify their being combined as a representation of a single construct (i.e., as being *unidimensional*). As Nunnally (1967, p. 251) has observed, “a test should ‘hang together’ in the sense that the items all correlate with one another. Otherwise, it makes little sense to add scores over items and speak of total scores as measuring any attribute.” To justify the combination of items in computing a respondent's overall or composite score on such a measure, the internal consistency of the item set must be established.

Some measurement instruments are developed to assess a multidimensional construct, like anxiety, or aggression, or perceived social support. Such instruments are designed specifically to dissect the structure of the constructs on which they are focused. This is perfectly legitimate. However, it is a mistake to use a multidimensional scale as if it were unidimensional by calculating a total scale internal consistency coefficient. Rather, the internal consistency of each of the scale's different dimensions (subscales) should be calculated independently. Simple summation of such scores is incorrect because it assumes implicitly that all subscales are measuring the same thing.

One of the earliest techniques to assess the internal consistency of a scale is a technique known as split-half reliability. **Split-half reliability** is assessed by randomly dividing a scale

into two sets containing an equal number of items, both administered to the same respondents, with a correlation calculated between these two summed scores.¹

If there is a high degree of interrelatedness among items, then the correlation between total scores from the scale's two halves should be strong, indicating that the items are focused on the same underlying construct (e.g., attitude or state). If the two halves of the measure do not “hang together,” this suggests that the scale items might not all be measuring the same underlying construct, which would result in a low correlation among the split halves of the test.

A more common alternative to the split-half technique is now used to determine a scale's internal consistency. This approach, **Cronbach's coefficient alpha**, is an index of the hypothetical value that would be obtained if all the items that could constitute a given measure were available and randomly combined to form every possible split-half measure of equal size (Cronbach, 1951). The average correlation between all possible pairs of these “split-half” tests is approximated by coefficient alpha.

Determining the alpha coefficient of a scale is relatively simple if one has a computer available (or if a relatively short test is being used). Computationally, we determine coefficient alpha as follows:

$$\text{Cronbach's } \alpha = \frac{k}{k-1} \left(1 - \frac{\sum \sigma_i^2}{\sigma_T^2} \right)$$

α = coefficient alpha, the estimate of whole-scale (internal consistency) reliability,

k = the number of items in the scale,

$\sum \sigma_i^2$ = the sum of the variances of each of the individual items,

σ_T^2 = the variance of the total scale.

Cronbach's alpha can range from .00 to 1.00. The degree of internal consistency usually is considered acceptable if this coefficient is .70 or better (Nunnally 1967), although the actual value depends on the extent of measurement error the investigator is willing to tolerate.²

From the internal consistency computational formula presented, we see that the *number* of items in a scale plays an important role in the scale's internal consistency reliability, as does the strength of the relations between items. If items are strongly related, alpha will be high. In addition, the formula suggests that all other things being equal, the more the items, the greater will be the scale's coefficient alpha. Thus, one tactic of enhancing alpha is to “lengthen” the scale – that is, to add items to it. *If* participants' responses to the new items are like their responses on the original set (that is, if the correlations between new and old items are high), the addition will enhance the coefficient of internal consistency (α).

The qualification presented at the beginning of the previous sentence suggests that considerable care be exercised when developing new items to add to an established set. This method of enhancing reliability is subject to the law of diminishing returns. Adding a reliable item to a five-item scale will have a greater effect on α than adding a reliable item to a 15-item scale. If the average correlation between items is reasonable (say, greater than .25), adding an item to a scale containing nine or ten items will have relatively little effect on α . Further, the higher the existing inter-item correlations, the less the effect of added items. A second qualifier involves the number of test items used. Adding items even remotely associated with the existing item set will improve alpha, but scales containing a high number of items become increasingly less impressive.

It sometimes happens that the coefficient of internal consistency is unsatisfactory even with relatively lengthy tests. One possible solution in situations such as these is to inspect relations among pairs of items, and to eliminate those items that do not correlate well with many of the

other items. A simpler method is to assess all item-total relations, that is, the correlation between participants' scores on each item and their total score over all items.³

If a specific item measures something different from that of the others in the item set, its relationship with the total score will be weak. This information will alert the scale developer that this item can be deleted and substituted with one that better represents the concept under investigation.

Our emphasis on internal consistency should not be taken to mean that all the items on a scale should be clones of one another. Ideally, the items of a scale should share a common focus, but should be different in other aspects that are irrelevant to this focus. For example, consider these two items, developed for a scale of attitudes regarding protection of the environment:

- The federal government should rule that automobiles must be developed so hydrocarbon emissions are completely reduced.
- All communities should employ both primary and secondary sewage treatment facilities before pumping wastes into public waters.

Agreement or disagreement with either of these items will be determined by several factors. The first item may be affected not only by concern for the protection of the environment but also by attitudes toward governmental intervention in private business, beliefs regarding the feasibility of complete elimination of hydrocarbon emissions, etc. Agreement with the second item will be affected by beliefs about the effectiveness of primary and secondary sewage treatment, the ecological integrity of the water supply, *in addition to* the central issue of the scale, attitudes regarding environmental protection. Both items potentially tap factors that are irrelevant to the issue of concern for the researcher, but these irrelevancies are *different* across items. Such heterogeneity of item content will produce some inconsistency of response (and hence, lower α), but if many such items are used, all focused on one common construct (though, perhaps, numerous noncommon determinants as well), the total item set will provide a better measure of the breadth or bandwidth of attitude than any single item. In this context, using the method of item-total correlations to examine relationships is useful in determining whether an item is a good representative of the construct under study, as the total score provides the best estimate of the construct.

It is possible for items on a test to be too closely interrelated (Cattell, 1972; Dawson et al., 1996). Consider the following example:

- Today's Volkswagen is a wonderful automobile.
- The new Volkswagen is a terrific car.

Clearly, we would expect the degree of relation between responses to these two items to be high. However, because they provide little, if any, non-redundant information, their consistency does not contribute much to the overall quality of the measure. This suggests that all things being equal, alpha coefficients in the high .90s should alert the researcher to item redundancy.

Temporal stability. Developing measurement scales having a high degree of interrelatedness among the items is one of the primary tasks of the test constructor. However, a second feature of reliability, called temporal stability, also merits close consideration in test construction. Questions pertaining to this aspect of scale quality are concerned with the degree to which the observations obtained on a test administration resemble those obtained on a second testing administration, which employs the same measure and the same respondents.

In considerations of temporal stability, researchers generally make use of one of two techniques. The most common is called **test-retest reliability** approach, which is assessed by administering a scale to a set of respondents, and later readministering it to the same respondents.

The purpose is to assess the association between scores obtained on the two administrations. Respondents' scores on the first administration are correlated to the scores they obtained on the second; a strong positive correlation is taken as evidence of (temporal stability) reliability.

An issue with the test-retest method concerns the optimal time lag between the two test administrations. Consider a few extreme examples: suppose an investigator employs a test-retest procedure with a delay of only 3 minutes between test administrations. Chances are good that the correlation between participants' scores on the tests would be nearly perfect. This would not, however, necessarily indicate that the scale would be consistent (i.e., replicable) if a longer time lag was used. A balance must be struck when deciding upon the appropriate interval between test administrations. Apparent temporal stability reliability will be enhanced artificially if participants can remember their previous responses and wish to appear consistent. Conversely, a very long delay between administrations can diminish temporal stability because people *do* change over time. Thus, even a very good test can appear unreliable if the temporal separation between administrations is extreme.

It is difficult to specify the "ideal" temporal lag between scale administrations with the test-retest method. Sometimes, even a modest time lag will artificially reduce the test-retest relationship. For example, suppose we constructed a scale to measure attitudes toward a well-known politician, and administered this scale to a sample of 500 respondents. One month later, we readminister the same scale to the same respondents. A 30-day lag does not seem overly long. However, suppose during that interval the politician was implicated in a juicy, well-publicized public scandal. These things have been known to happen in politics. In this case, the scale might not prove temporally stable. The attitudes of those who originally were favorably disposed toward the now-discredited public servant would be expected to change, whereas those of people who had always disliked the fallen politician would remain substantially the same. Such a change pattern would adversely affect the obtained test-retest correlation. In this case, differences in the observed scores would suggest unreliability. However, "history," rather than "instrumentation," is probably the real cause of the apparent lack of the measure's temporal stability.

A procedure was developed to circumvent some of the problems introduced by a temporal separation between test administrations and encountering identical items twice. Known as **parallel forms** (or alternate, or equivalent forms) **reliability**, it is assessed by devising two separate item sets intended to assess the same underlying construct, administered to the same participants at the same time. If a high correlation is obtained between scores on the two tests, it is interpreted as an indication of reliability of the instrument. Another way to verify that parallel scales were used is to show that the means and standard deviations of the two tests are similar if not identical. The rationale here is the same as that of the split-half approach, except that the two ("equivalent") forms are considered different versions of the same test.

These separate item sets represent two different questionnaire or instrument versions intended to assess the same underlying construct. For example, a researcher might wish to devise different versions of the same intelligence test, so that when many participants are administered the measure in the same room, the administration of version A and version B could be alternated by seating arrangement. Thus, having different versions of the same instrument would prevent participants from cheating off their immediate neighbors. Parallel forms reliability is classified under temporal stability because these two roughly equivalent forms (version A and version B) could have been created by the researcher at different time points.

As can be seen, questions of temporal stability can cause some difficulty for the test constructor. If a test is known to be temporally stable, then the explanation of changes between test administrations can be directed at external agents – e.g., a social intervention, an historical event, maturation, or other factors irrelevant to the measure itself. Such combined approaches should

not be avoided. In the ideal case, satisfying the full set of criteria of scale reliability enhances our confidence in our measure's temporal stability and internal consistency.

Cronbach and his colleagues proposed a more ambitious approach to reliability estimation, which they termed generalizability theory (Cronbach et al., 1972). This approach recognizes that irrelevancies (error) can be introduced into a test by many different factors, or *facets*, to use their terminology. These irrelevancies can reside in observers, items, contexts, occasions for measurements, respondents, etc. The more generalizable the instrument – that is, the more consistent it is across occasions, respondents, contexts, etc. – the better or more trustworthy is the instrument. Cronbach and his associates provided a framework for conceptualizing these different error sources, and for determining their individual impact on the measure under consideration. The generalizability approach has won adherents among some measurement experts (e.g., see Briesch et al., 2014; ten Hove et al., 2021; Vispoel et al., 2018), but is not yet widely adopted in practice. At a minimum, it provides a more comprehensive outline of the multitude of factors that may affect a score, and it alerts the researcher to the wide variety of factors that may affect the utility of an instrument.

Contemporary Measurement Theory

To this point, we have discussed the psychometrics of scale construction and assessment using classical test theory. The “classical” modifier is used to distinguish this traditional measurement framework from methods taking advantage of advances in contemporary or modern measurement theory, of which generalizability theory is but one. Contemporary approaches are designed to extend the fundamental ideas of classical measurement theory (e.g., Embretson, 1996; McNeish, 2018; Zickar & Broadfoot, 2009; Zimmerman, 2011).

Although no consensus exists regarding the features that constitute contemporary test theory, one definition is that it is characterized by the estimation of **latent factors** (Borsboom et al., 2003; Hancock & Mueller, 2011), a characteristic or construct not directly measured or observed that has removed measurement error (unreliability) from the factor. Latent factors are analogous to the “true” score component of classical test analysis. Measurement error becomes part of the scale score in classical testing, but its distorting effect is explicitly estimated and removed from a latent factor (DeShon, 1998). Latent factor analytic approaches may be divided into two general varieties (Borsboom et al., 2003). The first type includes factor analysis and structural equation modeling (the latter is elaborated in Chapter 10). The second type is based on item response theory and is discussed later in this chapter.

Contemporary testing approaches have been developed to address limitations of classical measurement approaches (Embretson, 1996; Vispoel et al., 2018). A major distinction in practice between the two approaches is the differential focus in the scale construction and evaluation process (Fan, 1998). Classical testing places great emphasis on the total scale, for example, the mean and reliability are calculated and reported to provide information about properties of the total scale. Contemporary test approaches, however, focus on individual items, offering greater information into how each item functions in the context of the total scale. Additional insights about item features help the scale developer evaluate each item's appropriateness for inclusion into the final measure or questionnaire.

Latent Factors and Correction for Unreliability

In classical testing, the observed score is viewed as containing not only the underlying true score but also measurement error. A scenario will help illustrate the measurement disadvantages

of a scale score contaminated by error. Before summing ten items to yield a summary score of altruism, we compute the internal consistency of the scale. Let's say this resultant composite is shown to be relatively reliable (Cronbach's $\alpha = .85$). Keep in mind, however, that the remaining .15 represents the unreliable remainder of the scale. This measurement error is a part of the summary score. Unfortunately, our inability to remove measurement error variance from the observed total score is the cost of using traditional classical testing. This contamination of an observed score is problematic when using the measure for hypothesis testing, or when using multiple measures, all of which contain their own unique error variance. When the altruism score is correlated to any criterion of interest, such as a score obtained from a volunteerism behavior scale, this statistical relation will be attenuated, its magnitude diluted because a proportion of each scale contains random error due to imperfect measurements. This obscures the detection of the real association between the two variables. A statistical adjustment proposed to remedy this problem of random error (unreliability) when performing statistical analyses using classic test theory is the *correction for attenuation* (Spearman, 1904), the formula for which is:

$$\text{Corrected } r = \frac{r_{xy}}{\sqrt{(r_{xx})(r_{yy})}}$$

r_{xy} = uncorrected correlation between measures X and Y,

r_{xx} = reliability of scale X,

r_{yy} = reliability of scale Y.

The correction for attenuation is an estimate of the correlation between two variables *if* the measures were perfectly reliable. As the reliability coefficients are input into the denominator, a scale lower in reliability will show a larger improvement after the correction, as it contains more random error. For example, suppose our results showed that the altruism and volunteerism scales correlated at .40, and that reliability for altruism is .70 and for volunteerism is .80. After accounting for the unreliability in the scale scores (i.e., correcting for attenuation), the deattenuated correlation jumps to .53. However, if the scales were perfectly reliable (i.e., $r_{xx} = r_{yy} = 1.0$), the "correction" for attenuation would not change the original result.

A larger correlation to account for the effect of unreliability is obtained because the correction now represents the correlation between two theoretically perfectly reliable factors (Bedian et al., 1997; Schmidt & Hunter, 1999). Despite being an important attempt to clean out random error, however, the correction for attenuation formula has not been widely adopted due to several criticisms (Charles, 2005). First, in some instances, it is possible to obtain out of bounds correlation coefficients ($r > 1.0$) after the adjustment. In other words, it is possible for the attenuation formula to *overcorrect* and inflate the estimate of the true relation, thus resulting in an impossible score. A second issue concerns the most appropriate type of reliability to be entered into the formula. Although parallel form reliability has been suggested as the best estimator, others have argued for internal consistency and test-retest approaches (Muchinsky, 1996).

To skirt these limitations, an alternative for isolating the replicable true score variance is to conduct analyses using latent factors (Bedian et al., 1997; de Shon, 1998). Latent approaches remove measurement error (unreliability) statistically – across the measured items and thereby extracts only the reliable or shared variance of the items into its factor. Applying the latent factors concept to our established understanding of classical test theory, only the reliable part of the variance reflecting the true score component remains, manifesting itself as the observed score. In other words, given that measurement error or unreliability has been statistically removed from the items, observed score = true score. (Again, true score exclusively pertains to

replicability of scores and implies absolutely nothing about whether the latent factor values of the measure of the construct are valid.) To fully appreciate the methodological appeal and usefulness of a latent factor, imagine calculating Cronbach's alpha on just the isolated latent factor. Theoretically, the alpha of the latent factor will be 1.00. Most accurately, to stipulate that the factor is perfectly reliable (or "error free," to use latent factor terminology) presumes that a representative sampling of items from the world of content was used to tap all aspects of the construct. It is prudent to avoid stating that the scale itself is perfectly reliable, instead saying that the latent factor has been corrected for unreliability for the items that were administered and is applicable only to that sample.

Factor Analysis

The earliest and best-known application of a latent technique is factor analysis. When scales are constructed with an emphasis on item variability, it can happen that participants' responses are so inconsistent over an entire set of items that few items correlate strongly with the total score, the summation of all the items in the scale. In this circumstance, it is important to determine whether the items, in fact, do access a single central, underlying (*unidimensional*) construct (or latent factor), or if the scale is *multidimensional*, that is, if it taps several related but not identical constructs. Using Cronbach's alpha or simply inspecting the pattern of item-total correlations is usually insufficient for evaluating the properties of a newly developed scale, especially if multiple dimensions or factors underlie the measure. To evaluate the possibility that the scale is multidimensional, the entire matrix of item intercorrelations can be entered into an exploratory factor analysis. This type of statistical analysis provides the researcher with information regarding the actual number of extracted factors or "subscales" that may exist in the instrument under construction, as evidenced by respondents' answers across all items. Based on this information, the investigator may decide to retain only a subset of the original items (e.g., those that form the most internally consistent subset of items, as indicated by a reliability analysis on the various subcomponents of the overall instrument), and to develop additional items to add to this subset in constructing an improved scale. The other items would be discarded, or used to create a separate scale or scales, with their own internal consistency coefficients, etc.

For example, suppose we developed and administered a set of items we believed tapped people's feelings about the preservation of the environment. If we were to perform factor analysis on the item set, we might find that one set of items that "hang together" (i.e., have high internal consistency) all have to do with participants' feelings of obligation to future generations. Another set of items might all have to do with the financial implications of environmental damage. These items do not relate much with the "future generations" items. In this case, we have two possible factors, or subscales, which measure different aspects of environmentalism worthy of study. We could add more items to each to create two revised and more reliable measures (recall that adding good items increases coefficient alpha) and re-administer both revised measures to a new sample. This procedure will enable us to determine the extent to which the new items hang together with the original set(s) of items. Eventually, an iterative process of factor analysis and reliability analysis will enable the investigator to generate a scale or subscales of acceptable internal consistency. The skill, insight, and experience of the scale constructor, the complexity of the issue under consideration, and the investigator's understanding of the issue and the respondent sample all play a role in the ultimate success of the measurement process.

An important and useful tool, factor analysis, may be divided into two categories (Bollen, 2002; Brown, 2006). Also known as common factor analysis, **exploratory factor analysis** is

a technique to determine the underlying factor structure of a set of items, if the researcher does not have prior hypotheses regarding the number of underlying factors and the items that might constitute each factor. The technique is data driven as the analysis offers a factor solution that optimally represents the underlying interrelations among items. **Confirmatory factor analysis** (CFA) is a technique to determine the goodness of the underlying factor structure if the researcher starts with hypotheses precisely stipulating the number of potential factors, which items should load on (or correlate with) which factors, and how the factors should be correlated. As the features of the factor structure are defined by the hypotheses, the resulting analysis reveals the degree to which items confirm the researcher's expectations (i.e., load on the prespecified factors). In exploratory and confirmatory factor analysis, the user interprets the pattern and strength of the relations revealed in the analysis.

Figure 3.1 depicts results of a confirmatory factor analysis, in which two latent factors are assessed by their respective items. Partitioned out of each measured item are two sources: the latent factor that produces or causes the variance of each item and its measurement error (E1, E2, E3, etc.) that explains the remaining variance of each item (Bedian et al., 1997). The factor loading or weight (interpreted like a correlation) between each item and the factor on which it was hypothesized to load also is shown in the diagram. Item 4 yields a poor factor loading that suggests that it does not “hang together” well with other items on that factor. A possibility is to allow the item to load on the other factor and then redo the analysis or delete the item if the modification does not yield a higher loading. The arrowheads of factor loadings always point from factor to items because the latent factor is presumed to drive its observable manifestations (measured items). This seemingly backward logic makes sense in theory and in practice: A factor loading indicates the extent a latent factor, statistically derived from the shared reliable or common variance across items, in turn explains or causes how participants respond to each item. For example, imagine that the items of a self-esteem questionnaire were subjected to factor analysis. The latent factor of participants' personal feelings of self-esteem is responsible for how they answer questions about self-esteem, consistent with the direction of the arrows from a latent factor to the measured variables. The opposite direction of arrows from the measured items to a latent factor would not make sense: Participants' responses and pencil markings on a self-esteem questionnaire are not responsible for producing their feelings of self-esteem. An item sharing more of its variability with other items (by more reliably “hanging together”) in

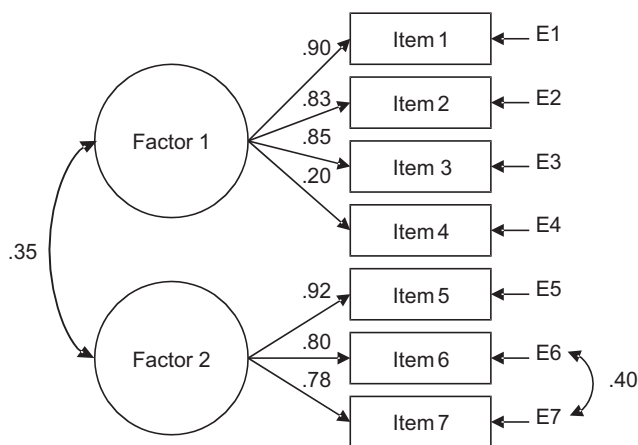


FIGURE 3.1 Results of a confirmatory factor analysis with two hypothesized latent factors.

the same factor will generate a larger factor loading and correspondingly is contaminated by less measurement error (unreliability).

In classical test theory, to compute an overall score, the researcher usually sums the items, a procedure that suggests item scores are additive (DeVellis, 2012). Doing so always assumes that items must have equal importance or equal factor loadings. Factor analysis allows computation of “factor scores,” in which items are given different weights based on their factor loading before being summed.

Also diagrammed in Figure 3.1, each measurement error term (E1 through E7) contains the variance portion of an item that is not explained by the latent factor. Measurement error may be decomposed into two elements (Bedian et al., 1997). Random error of each item (the unreliable portion of the item that does not “hang together” with the other items in that factor) has been separated from the latent factor, and therefore the factor has been corrected for unreliability attenuation. Furthermore, systematic error also has been estimated for each measurement error term, but only in instances when extraneous bias affects a single item (Brown, 2006). For example, if *only* item 2 was confusingly worded, and so the question was difficult for participants to comprehend, this source of systematic error would be extracted as part of E2. Because the bias was not shared among other items, it was statistically removed from the latent factor. Also in confirmatory factor analysis, it is possible to determine if systematic error is influencing more than a single item by testing whether two or more error terms show a non-negligible correlation (Shevlin et al., 2000). The estimated correlation between E6 and E7, for example, yields a value of .40, indicating that some source of systematic error that was not caused by the second latent factor is causing the unexplained variability of these two items in the same manner. Finally, the figure shows the positive correlation obtained between the two factors: higher scores on one factor are associated with higher scores on the other factor. Also, this demonstrates that although latent factors are not directly observable (they are observed through their items), they can be estimated and subjected to hypothesis testing.

Item Response Theory

Item response theory (IRT) is a technique to determine how each item operates in terms of difficulty, discrimination, and guessing with respect to the overall scale. Research applying IRT is dominated by scale development applications in educational testing and assessment. For this reason, the terminology of item response analysis makes use of language common in standardized testing, for the purpose of evaluating the appropriateness of items for inclusion into test banks (e.g., SAT, GRE, MCAT, LSAT, etc.). In the most widely used variety of IRT, the technique requires items using a binary response format (de Champlain, 2010). In educational testing, for example, responses to each item may be scored as either correct or incorrect. In other social sciences, attitudinal and behavioral questions might be dichotomized (recoded) to a yes/no or agree/disagree response.

The basic logic of IRT is to estimate how item responses in a test are related to performance on the test as a whole. The relation between the probability of item endorsement and performance or skill on its latent factor is evaluated. This is represented by an *item characteristic curve* or item response function, an S-shaped probabilistic curve that best approximates a model of the data. Figure 3.2 plots the levels of each of three hypothetical items in relation to gradations in its latent factor, which in IRT is known as *ability* or theta. Theta ranges on a continuum from low to high, ranking participants in terms of overall performance (or attitude or behavior) on the latent factor. On this metric, participants who performed at the average level relative to others on the scale have a theta of zero, those who performed better are represented by positive values,

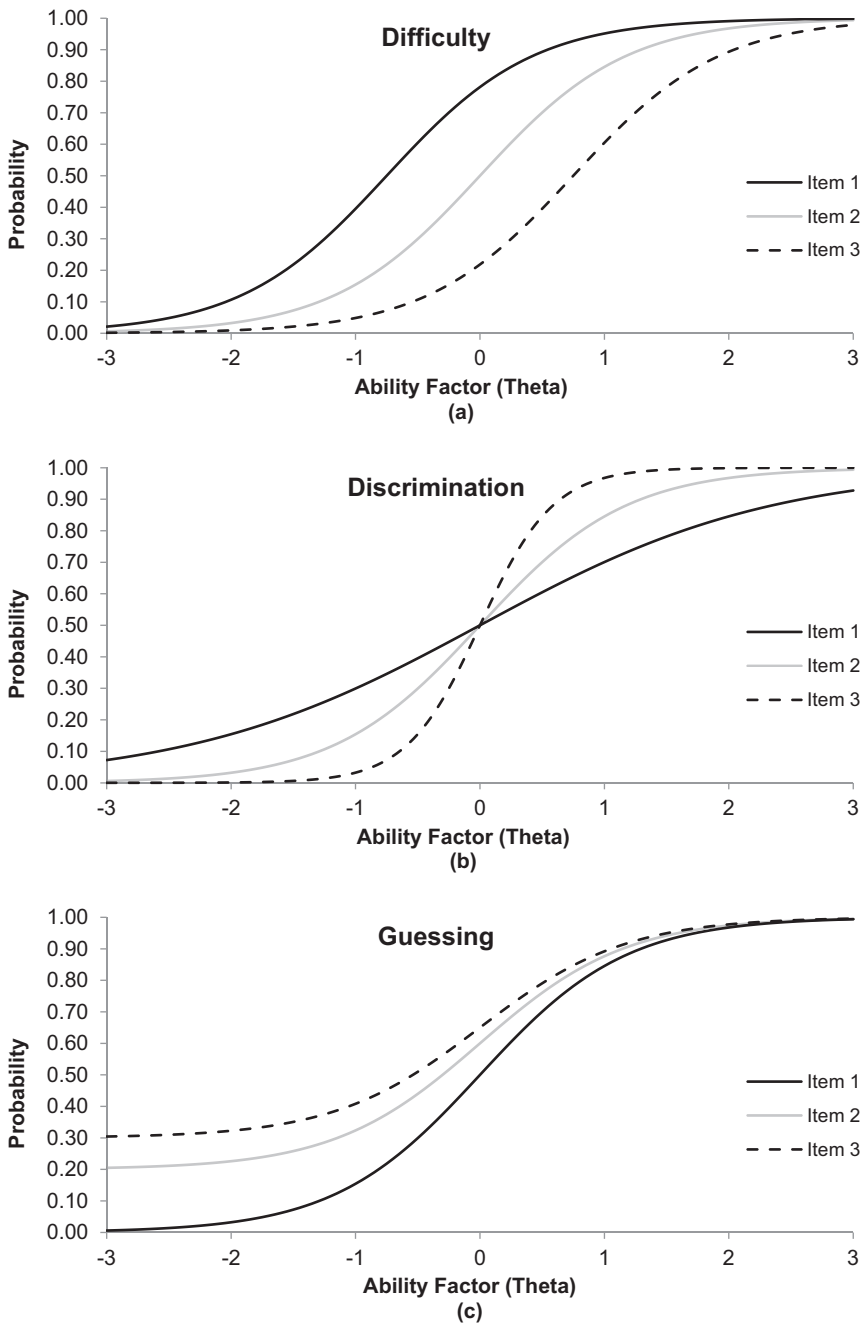


FIGURE 3.2 The three parameters of item characteristic curves in item response theory.

and those who performed worse have negative values at the left half of the ability continuum.⁴ Notice that the resulting characteristic curves are not straight lines, as would be found in the item-total correlation analysis of classical testing.

A response probability, ranging from 0.00 to 1.00, reflects the likelihood that an item will be answered correctly (endorsed), given any person's location on the latent factor. Probabilistic

language is used in IRT as there is no guarantee that a particular person will respond in a pre-determined way, but across respondents, an item characteristic curve estimates the probability of giving an affirmative response depending on where they are situated on the underlying factor. The concept of probability might be more intuitively grasped by thinking of the proportion or percentage of people in support of an item. An item with a .25 probability suggests that 25% of participants at that point on the latent factor endorsed or answered the item correctly.

To perform an item response analysis, it is crucial to satisfy two major assumptions about the data, or else results will produce inaccurate and uninformative estimates (DeMars, 2010; Henard, 2000). The *assumption of unidimensionality* requires that the factor be represented by items that assess no more than a single construct. Unlike factor analysis, IRT is inappropriate for analyzing a multidimensional item set. Only a single ability factor (or attitude) may be modeled in item response theory. If the scale is multidimensional, an imperfect solution is to select the dominant factor, usually the subscale containing the largest number of items in the inventory. Another possibility is to evaluate each subscale separately, but this fails to account for relations among the factors.

The *assumption of local independence* requires that all items be statistically independent. That is, a participant's response to an item should not depend on responses to any other items. The assumption is violated, for example, if the content of one item offers vital clues to answering another. The idea is that the factor should be the sole determinant of every item response on the test or questionnaire. This is the same assumption in other latent techniques: Only the underlying construct should drive responses to the manifest items.

Parameters of item characteristic curves. The relation between the ability factor and the probability of an item response is described using up to three parameters or characteristics (Coaley, 2010; de Ayala, 2009; DeMars, 2010). The three parameters are presented in Figure 3.2.

The *difficulty parameter* is the first property of an item characteristic curve and provides information on how likely people are to endorse or correctly answer an item relative to the other items. An item response analysis that estimates the difficulty parameter only is known as a Rasch model (Reid et al., 2007). An item deemed easy, or that has a high likelihood of a correct or positive response, is represented by an item characteristic curve situated in a location more to the left of the latent continuum. As shown in the top of Figure 3.2, a comparison of the three response curves reveals that item 1 is the easiest (or most endorsed), and item 3 is the most difficult (or least endorsed). For a person with a relatively low ability level, say theta of -1 , item 1 indicates a .40 probability, item 2 indicates a .15 probability, and item 3 indicates a .05 probability of endorsement. Conversely, for better performers who hold, say, a theta of 1 , the probability of positive endorsement for item 1 is .95, for item 2 is .85, and for item 3 is .60. Higher ability on the factor tends to correspond to an equal or higher rate of item endorsement. On either extreme of the latent factor, we find individuals who scored the lowest or highest on the total test, where each item has near 0 or 1.00 probability of being endorsed. In classical test theory, estimating the difficulty parameter is analogous to computing the proportion of participants endorsing each item and then comparing these proportions across items (Coaley, 2010; Sharkness & DeAngelo, 2011).

The *discrimination parameter* is the second property of the item characteristic curve. It offers information on how well an item successfully separates people of a particular ability on the construct. Discrimination is gauged by inspecting the slope for the item characteristic curve. An item possessing a steep slope that points more vertically is more effective in distinguishing respondents' performance on the factor. An item possessing a slope resembling a horizontal line is uninformative in discerning performance levels on the construct. In the middle of Figure 3.2b, item 1 possesses the flattest slope and item 3 has the steepest slope. From a relatively low ability

theta = -1) to high ability theta = 1, the curve for item 1 shows a respectable increase in proportion of endorsement from about .30 to .65, but item 3 shows a more substantial discrimination of abilities, from .03 to .95. This item comparison indicates that item 1 does not as strongly differentiate the ability of respondents and might not be a good candidate for inclusion in the scale. For example, a test item such as “2 + 2 =?” would be a poor item for a measure of intelligence because individuals at almost any level of intelligence would be able to answer it correctly. A psychometrically sound scale or test should show that each item successfully discriminates among its range of scores in terms of attitudes or ability. In classical test theory, discrimination is conceptualized as the degree of item-total correlation, and the magnitude of the correlation reflects the steepness of the slope (Coaley, 2010).

The *guessing parameter* is the third and final property of an item characteristic curve. It offers information about the likelihood that an item is vulnerable to being endorsed or answered correctly by anyone. Guessing may be understood as the probability of endorsing a particular item for people who had the lowest total performance on the scale. Imagine that students in your class received a surprise exam. Imagine further that some students did not spend any time hitting the books. The guessing parameter of the item characteristic curves will account for the fact that among students with the poorest total exam score, the responses of some items will tend to be correctly deduced even by those who did not study. To evaluate the guessing parameter, identify the probability of endorsement for an item characteristic curve located near the leftmost end of the factor (e.g., theta = -3). As diagrammed in Figure 3.2c, for people scoring poorest on the overall factor, item 1 has a near .00 probability of being guessed correctly, but the vulnerability for guessing correctly increases to .20 for item 2 and .30 for item 3. Applying these results to our previous example, students who did not study at all (and performed worst on the exam) are expected to show a 30% likelihood of correctly answering item 3. Guessing plays virtually no role in obtaining a correct response for item 1. It is practical to eliminate items showing a high probability of being guessed correctly by respondents.

Simultaneously estimating all three parameters. The collective interpretation of these three parameters using item response theory is illustrated using a sample of alcohol users. The data for this example were drawn from an IRT study that examined patterns of alcohol consumption throughout the week (Lac et al., 2016). Respondents answered yes/no to questions regarding whether they usually drink alcohol on each of the five weekdays (e.g., “typical Monday”). The five items were used in an item response analysis, with results shown in Figure 3.3. In this scenario, the ability factor represents a continuum of the total number of weekdays that respondents usually consumed alcohol. That is, lower ability denotes people who infrequently engage in weekday drinking, and higher ability denotes those who drink on more of these days. Evaluating the difficulty parameter, the item response curve for Friday is positioned near the left end of the factor and is therefore the most common drinking day. Specifically, people low on drinking ability (e.g., theta = -1) are most likely to drink on Friday. The implication is that for those who are least likely to drink alcohol across the weekdays, there is a small, near zero probability of consumption from Monday to Thursday, but these same individuals have a greater likelihood of consuming on Friday, the start of the weekend. In contrast, heavier daily drinkers (e.g., theta = 1) have a higher probability of alcohol use *not only* on Friday but also on the other weekdays, with the lowest probability on Tuesday.

The success of each weekday in discriminating along the continuum of overall weekday drinking is evaluated with the discrimination parameter. Among people lower on daily weekday drinking (e.g., theta = -1), Friday has the steepest slope in effectively discriminating such participants. However, among people high on daily weekday drinking (e.g., theta = 1), Friday does not offer as much discernment compared to the other weekdays. The slopes for the Monday

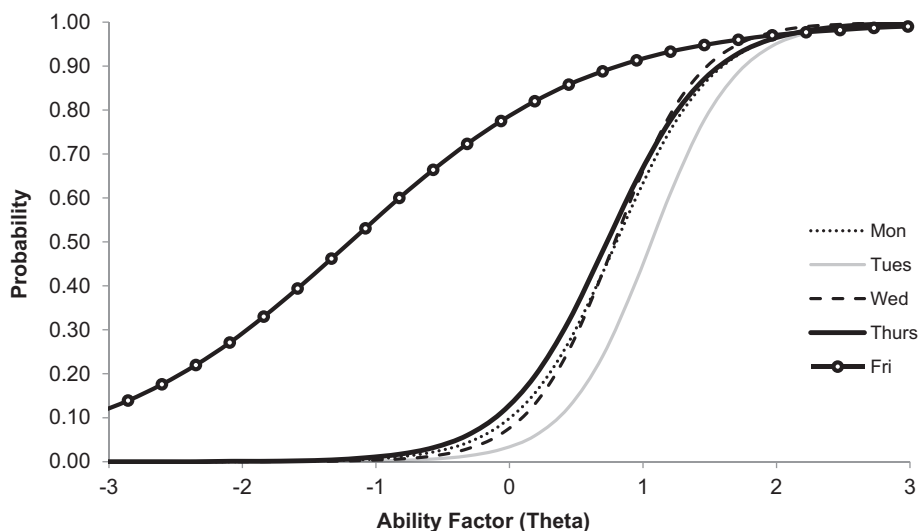


FIGURE 3.3 Item response theory analysis of weekday alcohol consumption.

to Thursday items are almost flat at the low ability range, signifying that these days offer very little information for discriminating individuals who do not drink much. In contrast, Mondays to Thursdays better differentiate those who are relatively higher on the latent factor of daily weekday consumption.

Finally, consider the guessing parameter: The probability of drinking on each of the five days for respondents who drink the least or very minimally. For a person at the extremely low end of the drinking ability parameter (e.g., $\theta = -3$), the probability of guessing is near .00 for Monday to Thursday, but approximately .13 for Friday. In other words, even among infrequent weekday drinkers, 13% of them will succumb to temptation when Friday arrives.

Information obtained in item response theory analyses could be used in various ways, depending on the purpose of a study. It appears that Friday and Tuesday drinking least resemble the response patterns for the other three days; consequently, a summary score using only items representing the other three weekdays would produce the most reliable scale. For a different purpose, the item response curves for Monday, Wednesday, and Thursday were found to be very similar for the difficulty, discrimination, and guessing parameters and therefore provide highly redundant information. Thus, if you knew whether someone consumed alcohol on Monday, you would likely be able to predict whether that person is also drinking alcohol on Wednesday and Thursday, because these three items exhibit almost identical item characteristic curves.

A practical advantage of item response analysis is that test developers can predict the ability level of respondents who have completed only a few items selected from the entire inventory, test, or questionnaire (Reid et al., 2007). This time-saving application is found in computerized adaptive testing, in which new test takers are administered a relatively small subset of items from a large test bank (Hol et al., 2008). The sequence of items received by a participant is a function of the test taker's competence (or ability, if using IRT language) on the previous items. Thus, the test adapts to how well a person is performing at any point during the examination. A person deemed to possess low ability, due to poor performance early on the test, will continue to receive "easy" questions that previous test administrations have established as having a high rate of being answered correctly by most people. A person showing high ability on early items will

continue receiving more “difficult” questions that have a lower probability of being answered correctly by most people. Because adaptive testing does not require the administration of the entire pool of items, it is an efficient approach for isolating and narrowing a person’s performance during an exam. The person’s overall score is calculated by locating or linking where he or she falls on the ability continuum in comparison to others whose scores were used to establish the item response curve norms. Conversely, just knowing a person’s ability (theta) level makes it possible to form inferences about the probability of correctly answering an item a respondent never received.

Conclusion

Our conceptualization of a phenomenon or entity consists of diverse operations caused by unobservable underlying factors. When you are asked to describe “love,” this abstract concept may conjure up love of your romantic partner, love of your parents, or even love of your dog. You do not go around thinking that your gut feeling of love is impure and contaminated by measurement noise. Latent factors also allow items to possess appropriate weights in relation to their adequacy in tapping the aspects of the attribute. In a sample of newlywed participants, a factor analysis of the data might show that the largest factor loading or weight arises from the item of “romantic or passionate love,” as it more quintessentially captures how this cohort perceives the emergent quality of love. Equal weights used in non-latent approaches presume that the different operations reflecting love (e.g., intimacy, passion, commitment) are equally sufficient in capturing the replicable and unobservable essence of love.

The use of latent factors in techniques applying modern test theory can be understood as an extension of classical test theory. Some clear benefits emerge from the use of latent factors, but this is not to imply that there are no advantages to classical test analysis, as the primary benefits include ease of learning, estimation, and interpretation. Thus, classical test analysis remains a popular method for evaluating the quality of measures. Comparative studies have discovered that when both non-latent and latent techniques are undertaken and contrasted on the same scale or test, they produce highly parallel results (Fan, 1998; Sharkness, 2011). These psychometric models should be viewed as complementary, rather than entirely different ways of conducting research. If possible, the researcher may report results of classical test analyses (e.g., item-total correlations and overall mean) as well as latent analyses (e.g., factor analysis or item characteristic curves showing relations between each item and the latent factor) in the same study, to allow the audience to make better-informed conclusions about the utility of the measures to be used in examining the substantive research question.

Notes

- 1 Various means are used to split the total item set in half: even numbered items are contrasted with odd numbered items, the first half compared with the second half (probably a poor choice because of possible fatigue effects on the part of participants), the total item set is randomly split into two groups, etc.
- 2 Cronbach’s alpha is a generalization of the KR-20, an index of internal consistency for scales using dichotomous items (Kuder & Richardson, 1937). For scales using binary items, the formulas for Cronbach’s alpha and the KR-20 yield identical values.
- 3 It is good practice to adjust these “item-total correlations” statistically, so as to remove the influence of the particular item under consideration on the total score. This adjustment becomes especially important when the scale is composed of relatively few items, because in such cases, the contribution of any given item to the total score is great.

References

- American Psychological Association (1985). *Standards for educational and psychological testing*. Washington, DC: APA Press.
- Bedian, A. G., Day, D. V., & Kelloway, E. K. (1997). Correcting for measurement error attenuation in structural equation models: Some important reminders. *Educational and Psychological Measurement*, 57, 785–799.
- Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, 53, 605–634.
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, 110, 203–219.
- Briesch, A. M., Swaminathan, H., Welsh, M., & Chafouleas, S. M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of School Psychology*, 52(1), 13–35.
- Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.
- Cattell, R. B. (1972). *Personality and mood by questionnaire*. San Francisco, CA: Jossey-Bass.
- Charles, E. P. (2005). The correction for attenuation due to measurement error: Clarifying concepts and creating confidence sets. *Psychological Methods*, 10, 206.
- Coaley, K. (2010). *An introduction to psychological assessment and psychometrics*. Thousand Oaks, CA: Sage.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.
- Dawson, E. J., Crano, W. D., & Burgoon, M. (1996). Refining the meaning and measurement of acculturation: Revisiting a novel methodological approach. *International Journal of Intercultural Relations*, 20, 97–114.
- de Ayala, R. J. (2009). *Theory and practice of item response theory*. New York: Guilford.
- de Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education*, 44, 109–117.
- DeMars, C. (2010). *Item response theory*. New York: Oxford University Press.
- DeShon, R. P. (1998). A cautionary note on measurement error corrections in structural equation models. *Psychological Methods*, 4, 412–423.
- DeVellis, R. B. (2012). *Scale development: Theory and applications* (3rd ed.). Thousand Oaks, CA: Sage.
- Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, 4, 341–349.
- Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, 58, 357–381.
- Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.
- Hancock, G. R., & Mueller, R. O. (2011). The reliability paradox in assessing structural relations within covariance structure models. *Educational and Psychological Measurement*, 71(2), 306–324.
- Henard, D. J. (2000). Item response theory. In L.W. Grimm & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 67–97). Washington, DC: American Psychological Association.
- Hol, A. M., Vorst, H. C. M., & Mellenbergh, G. J. (2008). Computerized adaptive testing of personality traits. *Journal of Psychology*, 216, 12–21.
- Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2, 151–160.
- Lac, A., Handren, L., & Crano, W. D. (2016). Conceptualizing and measuring weekend versus weekday alcohol use: Item response theory and confirmatory factor analysis. *Prevention Science*, 17, 872–881.
- McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, 23, 412–433.
- Muchinsky, P. M. (1996). The correction for attenuation. *Educational and Psychological Measurement*, 56, 63–75.
- Nunnally, J. C. (1967). *Psychometric theory*. New York, NY: McGraw-Hill.
- Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York, NY: Routledge.
- Reid, C. A., Kolakowsky-Hayner, S. A., Lewis, A. N., & Armstrong, A. J. (2007). Modern psychometric methodology: Applications of item response theory. *Rehabilitation Counseling Bulletin*, 50, 177–188.
- Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence*, 27, 183–198.

- Sharkness, J., & DeAngelo, L. (2011). Measuring student involvement: A comparison of classical test theory and item response theory in the construction of scales from student surveys. *Research in Higher Education*, 52, 480–507.
- Shevlin, M., Miles, J. N. V., Davies, M. N. O., & Walker, D. S. (2000). Coefficient alpha: A useful indicator of reliability? *Personality and Individual Differences*, 28, 229–237.
- Spearman, C. (1904). Proof and measurement of the association between two things. *American Journal of Psychology*, 15, 72–101.
- ten Hove, D., Jorgensen, T. D., & van der Ark, L. A. (2021). Interrater reliability for multilevel data: A generalizability theory approach. *Psychological Methods*, 27, 650–666.
- Vispoel, W. P., Morris, C. A., Klinc, M. (2018). Using generalizability theory to disattenuate correlation coefficients for multiple sources of measurement error. *Multivariate Behavioral Research*, 53, 481–501.
- Zickar, M. J., & Broadfoot, A. A. (2009). The partial revival of a dead horse? Comparing classical test theory and item response theory. 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. 37–59). New York, NY: Routledge/Taylor & Francis Group.
- Zimmerman, D. W. (2011). Sampling variability and axioms of classical test theory. *Journal of Educational and Behavioral Statistics*, 36, 586–615.

4

MEASUREMENT VALIDITY

Developing measures that meet the criteria of reliability in terms of internal consistency, temporal stability, or inter-observer agreement satisfies a basic requirement for the operationalization phase of a scientific investigation. In the first chapter of this text, however, we made a distinction between the adequacy of an operation in terms of replicability and its adequacy as a manifestation of a theoretical construct. This distinction marks the difference between the *reliability* of a measuring device and its *validity*. Whereas reliability has to do with the consistency of measurement, the validation of operations relative to the hypothetical concepts under investigation (constructs or latent factors) is crucial from the standpoint of theory development.

The procedures usually followed to generate reliable scales of individual differences (attitudes, abilities) could lead to internally consistent and temporally stable instruments that had no relation whatsoever to the theoretical attributes that motivated the research in the first place. Consistency of responses, from item to item, time to time, or observer to observer, although necessary, does not establish measurement validity. Although some degree of response consistency is essential in diagnosing any underlying attribute, the validity of a measuring instrument must be studied through use of operations beyond those applied to assess reliability.

Measurement validity refers to the extent of correspondence between variations in the scores on the instrument and variation among respondents (or other objects of measurement) on the underlying construct being studied. In classical test theory, the true score represents a measure of the replicable “shared variation” or the “common factor” that underlies participants’ responses to all items. Whether this response factor adequately reflects the specific conceptualization the investigator wants to measure, however, is still a matter for investigation. It is important to keep this point in mind when validating any measure. Validation always requires empirical research beyond those used in instrument development and refinement. This validation process focuses on the relation of the scale with some other indicators of the construct being studied.

Nunnally (1967) noted that validity is a relative, descriptive term, not an all-or-none property, and his view accords with Messick’s (1989, p. 13) definition of validity as a “judgment of the degree to which evidence and theoretical rationales support the *adequacy* and *appropriateness*” of a construct. This definition suggests that validity is not a thing or a feature of a measure, but rather an aspect of the interpretation of a measure. As such, validity is always open to question,

review, and revision – it is never a closed issue, but rather a continuous process. It is “the best available approximation to the truth or falsity of propositions” (Cook & Campbell, 1979, p. 37), but the judgment must always be tempered by recognition that the best available evidence may change, and thus the evaluation of validity of a construct also may change.

Types of Measurement Validity

With this general introduction to what the concept of validity is, and what it is not, we consider some specific subcategories of validity of central concern to social scientists. An appreciation of these more specific features of validity provides a more complete understanding of validity as it is applied in the scaling of stimuli or people, which are considered in Chapters 16 and 17, respectively. In discussing these various forms of validity, it is important to keep in mind that they all are component parts of what we term **measurement construct validity**, which is concerned with the validity of the measure with respect to the theoretical construct of interest. The various components of construct validity, presented in the following paragraphs, are important insofar as they reflect the underlying theoretical construct that the measure is intended to assess.

Face Validity

Face validity is based on informed impressions regarding the extent that a measure appears to capture a construct. A scale designed to assess interpersonal satisfaction probably should not contain the question, “Do you enjoy the company of cats?”, but it might include “Are you happy with your interpersonal relationships?” Claims of face validity are not usually based on systematic established procedures that characterize other forms of validity assessment (deVellis, 2012). For this reason, many do not embrace face validity as a legitimate form of validity.

An important consideration when considering face validity is that it sometimes might be a good thing to avoid, especially if the goal is to conceal the purpose of the question from respondents. Measures that have high face value in tapping socially unacceptable behaviors or social stigmas might deter honest responses. The Minnesota Multiphasic Personality Inventory (MMPI), designed to assess psychopathology, contains questions purposely designed as being low on face validity (Gynther et al., 1979). An example of an MMPI item requiring a true/false response is, “I prefer a shower to a bath” (Ben-Porath, 2012). The question appears to assess hygiene preferences quite innocuously, but people who would rather sit than stand in warm water tend to score lower on empathy and compassion for others. Although initially selected because of low face validity, inclusion into the final version of the MMPI occurred only after other types of measurement validity established the item as an acceptable indicator of the construct (empathy). Thus, the mere fact that a scale’s items do not seem reasonably related to the construct under study (i.e., does not have “face validity”) is not a sufficient reason to exclude them. But neither is it sufficient to justify an item’s inclusion as a valid indicator in the absence of other corroborating evidence, no matter how good it looks (face validity). Like most other measurement issues, a confluence of different methods may prove the best road on the way to developing valid measures.

Content Validity

The extent a measure adequately represents (or samples) the complete range or breadth of a construct under consideration is termed **content validity**. This form of validation is used primarily

in instances involving objective assessments of knowledge. For example, in establishing the content validity of a test of eighth-grade arithmetic skill, we would be concerned with whether the scale adequately sampled the range of mathematical skills that an average eighth grader should know. If the test focused exclusively on addition and subtraction, it is obvious that it would not have sampled enough of the hypothetical domain, and some modifications to include a greater variety of items (e.g., multiplication and division) would be indicated.

With subjective measures of knowledge or ability, constructing measures with adequate content validity is not overly difficult if the domain of interest is relatively well specified. When dealing with issues of opinion rather than facts, the situation is not so clear-cut. In such cases assessment of content validity often is a subjective operation. To help ensure content validity, many researchers begin the scale construction process by generating large numbers of diverse items. Whether this represents an adequate sampling of content is based almost solely on the subjective judgment of the researcher constructing the scale, and as such is subject to bias. Three systematic procedures offer ways to evaluate content validity.

The first approach requires a thorough literature review of the topic to acquire insights about the appropriate types of items or factors to ensure a fair and comprehensive coverage of the construct. Examining concepts and definitions commonly employed in the critical research area will help with decisions regarding content breadth and also how best to phrase these items. A second means of establishing content validity is by using experts, who provide advice regarding the adequacy of a scale's coverage. Although far from infallible, this usually provides more trustworthy information than that of an investigator working independently. The third and most empirically based way to evaluate content validity is to use factor analysis to assess the adequacy with which various features of the construct are represented in the measure (Bryant, 2000). Doing so will reveal not only the number of factors or content areas emerging from the scale, but also the number of items that tap each factor. Factors showing inadequate coverage probably will be represented by fewer scale items relative to others. Given this information, the researcher might revise the scale by devising additional questions for the underrepresented domains.

Criterion Validity

The procedures designed to assess the validity of a scale should vary according to the purposes for which the instrument is devised. **Criterion validity** is concerned with the extent that a measure is related to, or explains, a target outcome or criterion, usually a behavior. For example, a scale constructed to predict likely voter choices in a presidential election may use the respondents' actual vote as a criterion. Depending on the distance in time between administration of the scale and the outcome, criterion validity may be divided into two types (Bryant, 2000): **Concurrent validity** assesses the extent that a measure is related to, or explains, a relevant criterion behavior, with both variables assessed at the same occasion. An issue with concurrent validity is that it is impossible to unravel the temporal precedence of events. For example, a multiple-item scale of exercising attitudes was developed and then used to explain the outcome of number of times exercising. Did positive attitudes about exercising explain frequent exercising behaviors, or does frequent exercise create positive attitudes toward such behaviors?

Predictive validity is concerned with determining the extent that a measure is related to a relevant criterion behavior assessed on a subsequent occasion. It is considered the more desirable and rigorous of the two forms of criterion validity.

Predictive validity is of major concern when the purpose of a measure is to anticipate either the likelihood or the extremity of some behavior or outcome of interest occurring at a later time. The behavior itself serves as the criterion, and the strength with which the scale predicts

the level of the criterion is taken as an indication of the scale's predictive validity. It resembles concurrent validity except its prediction involves a temporal lag between measure and outcome. A test of reading readiness, for example, should enable an investigator to discriminate the "ready" from the "unready," to predict which children are most likely to succeed in grasping the fundamentals of the essential skill. To assess the degree that the test successfully discriminates this behavior, a researcher could administer the test (the predictor) to preschoolers on the first day of school and correlate these results with the scores received on a standardized test of reading achievement after their first year of formal schooling. The strength of the relation would suggest the predictive validity or invalidity of the test.

In judging the utility of prediction to establish the validity of a measure, three important limitations should be kept in mind. First, as other variables may influence the magnitude of a relation, predictive validity in and of itself is not sufficient to confirm or to refute the validity of a measure. A failure to obtain a strong relation in the foregoing example might have been due to a massive failure of the instructional system to teach reading effectively to the children in the sample. Or, it might have indicated that the standardized test was not appropriate for the sample.

A second, more practical limitation on the usefulness of prediction in establishing the validity of a scale has to do with the relative absence of suitable criteria. Few of the variables of social science are as easily quantified as standardized reading achievement. Typically, our investigations, and consequently, our measures, are constructed to assess complex, abstract, qualities. In these instances, predictive validation approaches often are not fully useful, because appropriate criteria against which predictions might be compared do not exist – they must be inferred. For example, suppose we wanted to develop a measure of religiosity. In this instance, predictive validity would be difficult to assess. What is an acceptable criterion measure? Attending religious services? Helping those in need? Being kind to strangers? Any of these possibilities might characterize people of either high or low religiosity. One might argue, for example, that compassion is not necessarily indicative of religiosity. Nonreligious people, that is, can be compassionate or not, just as religious people can.

Convergent Validity

If a measure is a valid indicator of a construct, it should be correlated positively with other types of measures of the same or similar constructs. The process of testing this form of validity is termed **convergent validation**, which assesses the extent that measures of theoretically related constructs are related empirically. Convergent validity should not be confused with concurrent validity described earlier. The other measures used to assess convergent validity may be scales of theoretically similar types of constructs, but also could involve measures of the same construct employing different methods, such as peer reports, coded observational records, or other types of measurement formats.

Convergent validation subsumes other validity forms – predictive validity, concurrent validity, and so on. In the ideal case, indicators that are as dissimilar as possible are used to estimate the convergent validity of a measure. This view distinguishes the search for convergent validity from the assessment of reliability, in which maximally similar measures are sought. In the quest for convergent validity, measures as dissimilar as possible are developed and used so as to increase the likelihood that they share different sources of error. The more similar the measures, the greater the probability that they are prone to the same irrelevancies, thereby compounding error (Crano, 2000, p. 40).

Let's return to our earlier example, which involved developing a measure of religiosity. One might expect this measure theoretically to be related to other measures with which religiosity is

thought to covary. For example, we might expect this measure, if it is valid, to relate to similar constructs, such as measures of compassionate beliefs or conservative ideology as informed by previous studies. If these expected relations do not materialize, two possibilities arise: the most obvious is that our scale does not adequately assess the construct (it is assumed that the instruments measuring the related constructs, compassion and conservative ideology, are relatively valid). The second possibility is that the scale is valid, but that the theory on which the hypothesized relations are based is incorrect, and thus no relation between the constructs should be expected. However, if another scale of religiosity does exhibit the hypothesized relation pattern with the external validating scales (compassionate beliefs and conservative ideology), the investigator is left with the first alternative – the new measure probably is invalid and in need of reconceptualization.

Of the many ways in which convergent validity is assessed, perhaps the most obvious is the **known groups method**. In this validation approach, a measure designed to assess a proposed construct is given to different groups of people who are known to differ on an attribute that is related to the focus of the instrument. If the scale measures what it purports to, then these groups should have significantly different scores, and these differences should be known in advance of the test's administration. For example, if an investigator has devised a scale to measure discrimination against Black people then the average scale score of a mob of Ku Klux Klan members should be different from that of a group of American Civil Liberties Union members. The known groups method is considered a type of convergent validity because the new measure of perceived discrimination is expected to be correlated with the grouping variable. If no differences are found, or if the differences are in a direction opposite to that expected, the validation process has failed.

The known groups method has been used in several different situations, and generally proves a convincing technique in establishing the case for validity. For example, Wheeler and Austin (2000) were interested in validating a scale to assess adolescent women's grief responses. They assembled a sample of young adolescent women, 13–19 years of age, some of whom had recently suffered a perinatal loss. As indicated by the authors' "Loss Response List," those women who had lost their babies exhibited significantly more grief and depression than the women who had not endured such a loss. These score differences were as might be expected based on theory and common sense, thereby supporting the validity of the measure.

Although the known groups method is useful, there are some contexts that preclude its use for practical purposes. Sometimes, for example, the appropriate groups will not cooperate. Sometimes, the quality under consideration does not suggest identifiable groups who would be expected to differ greatly on the construct. As such, the application of the known groups technique is limited by the nature of the underlying construct the instrument is designed to measure. In some areas of research, alternative procedures must be employed to establish convergent validity by relying on correlations among related measures.

Discriminant Validity

In the convergent validation process, we generate a series of hypotheses regarding probable relation expected between the measure of the construct under development and other, more established scales of related constructs, independent of extraneous factors (e.g., similar methods of measurement). In a complementary analytic approach, we also can assess validity by examining the extent to which the construct of interest is distinct from other measures with which it should be *unrelated*. **Discriminant validity** is the extent that measures of constructs that are theoretically unrelated are empirically unrelated. The religiosity scale should not overlap highly with measures of patriotism or intelligence.

TABLE 4.1 Developing and validating a “Student Research Potential Scale.” Examining the major forms of measurement validity

<i>Measurement validity</i>	<i>Example process</i>
Content Validity	The published literature is scoured to review the characteristics and profiles of students who are research productive. Professors prolific in research are interviewed to identify the most common beliefs and views about why they and their students enjoy research. Two common themes are identified: (a) Creativity in generating novel research ideas, and (b) Passion for learning about research methods. Diverse items are generated based on these domains.
Criterion Validity	After administering the scale of items to college students, behavioral outcomes are completed by the same sample one year later to assess predictive validity. These criterion indicators include involvement in research projects with faculty and number of research conferences attended during the one-year span. The Student Research Potential Scale should longitudinally predict or explain variability in the behavioral measures.
Convergent Validity	Similar, but established, measures in the literature are administered along with our scale: The Need for Achievement Scale, Academic Aspirations Scale, and Workaholic Measure. The Research Potential Scale should reveal relations to these other measures: All stem from a strong work ethic. To demonstrate why this new scale offers non-redundant information from these other similar scales, relations should not be strong.
Discriminant Validity	Scales unrelated to our new construct are also administered alongside our scale: Need for Security Inventory, College Partying Scale, and Clown Phobia Scale. Relations should be negligible, if not zero, to show the Student Research Potential Scale is tapping a phenomenon different from these other assessments.

To help illustrate the various forms of validity that we have considered thus far, Table 4.1 describes the procedures associated with different types of measurement validity using a hypothetical example that involves developing a scale of Student Research Potential. As shown, content validity is enhanced through a thorough review of the literature and intensive interviews of prolific student researchers to isolate attitudes and behaviors that seem to be associated with research productivity. After developing a reasonable number of items designed to tap this construct, the researcher administers these items to a large sample of college students, and one year later assesses their research performance through measures of the number and quality of research projects in which they have been involved, the number of scientific conferences they have attended, the number and grades on research-based courses they have taken, and so on. If students' scores on our new scale are correlated with these indicators of potential research productivity, the scale will be judged to have good predictive validity, a component of criterion validity.

While indicators of criterion validity are collected, the researcher also could administer established tests of features that *theoretically* should be associated with research potential. If these measures relate to (or converge with) our measure in expected ways, the new scale is said to have convergent validity. Finally, the researcher also might have administered scales with which our new measure would *not* be expected to relate. This is done to ensure that the new measure discriminates among constructs – that is, that it relates to measures with which theoretically it should relate, and that it does not relate to measures with which it should not. If this latter criterion is satisfied, the Student Research Potential scale is said to have discriminant validity.

The Multitrait-Multimethod Matrix: Combining Convergent and Discriminant Validity

In general, construct validation is an approach whose aim is to establish the reality of a psychological concept – it is a test of whether the hypothesized construction plausibly exists. If it does, then it should demonstrate predictable patterns of relations with other constructs. When attempting to validate a construct, investigators must develop a set of hypotheses regarding other constructs with which their measure should or should not relate. If a measure is to be judged a valid indicator of the construct, the hypothesized relations should exist, and measures purportedly assessing these associated constructs should be related.

Operations that assess the relations between a new measure and other established measures with which the new test was thought to relate were termed *convergent validation techniques* by Campbell and Fiske (1959), because in essence the measures converge upon, or define, a hypothesized network of interconnected traits, processes, dispositions, and behaviors. A successful convergent validation operation not only suggests that the critical scale is an adequate measure of the construct in question, but also bolsters the theoretical position that was used to develop the hypothesized relations that formed the basis of the validation process.

Discriminant validation techniques also can be used in conjunction with studies of convergent validity to assess scale quality. Both convergent and discriminant validity are subclasses of the more general concept of construct validity and are investigated simultaneously in the multitrait-multimethod matrix approach developed by Campbell and Fiske (1959). Each of these two validity forms provides different and useful information about the quality of an instrument.

In the **multitrait-multimethod matrix** (MTMM) approach, multiple measures are used to assess the association of theoretically related, but different constructs, over and above the association that might come about simply from having shared the same method of measurement. The MTMM is a logical extension of the complementary principles of multiple operationalization and triangulation introduced in the first chapter of this text. The technique involves computing a correlation matrix that reveals the relations among a set of carefully selected measures. The measures are chosen to represent a combination of several different (theoretically relevant) constructs, each assessed by several different methods of measurement. Each construct is measured by *each* measurement technique. Analysis focuses on the relations between measures theorized to assess the *same* construct. These relations are compared with those involving measures of *different* constructs that happen to be measured by the same measurement techniques. The pattern of interrelations among traits sharing the same and different methods of measurement helps us determine the construct validity of the measures under study.

A concrete example may help illuminate this approach. Suppose a social developmentalist wanted to create a new scale of aggressiveness for young children using the method of teacher's behavioral observation. Based on theory, the researcher believes aggressiveness scores should be distinct from sociability and intelligence. To use the MTMM approach, the researcher finds two previously established measures of aggressiveness shown in the prior research to possess desirable psychometric properties. The first of these scales uses the method of self-report to measure aggressiveness, and each child completes this scale individually. The second scale employs the method of peer ratings, in which the best friend of each child is asked to rate the respective child's aggressiveness. The new measure, remember, uses the method of behavioral observations, in which an observer (in this case, the child's teacher) observes the playground interactions of each child for 10 minutes and uses a checklist to quantify their aggressiveness. Then, the researcher finds *established* scales of both sociability and intelligence, each of which is assessed by the same three methods used to assess self-esteem (self-ratings, peer ratings, and teachers'

TABLE 4.2 A 3 (Methods: A, B, C) × 3 (Traits: 1, 2, 3) multi-trait multimethod matrix

Traits	Method A (Self-Rating)			Method B (Peer Rating)			Method C (Teacher's Behavioral rating)		
	1 (Self-Esteem)	2 (Sociability)	3 (Aggression)	1 (Self-Esteem)	2 (Sociability)	3 (Aggression)	1 (Self-Esteem)	2 (Sociability)	3 (Aggression)
Method A (Self-Rating)	(.90)	.50	.40						
		(.88)	.30						
			(.83)						
Method B (Peer Rating)	.70	.11	.10	(.92)	.42	.39			
		.20	.09		(.91)	.31			
			.65			(.83)			
Method C (Teacher's Behavioral rating)	.67	.23	.18	.69	.44	.34	(.93)	.31	.28
		.20	.12		.41	.28		(.89)	.21
			.58		.29	.61			(.80)

Note: Reliabilities are the parenthesized values on the main diagonal. Monotrait-monomethod (Validity) values are in bold type. Heterotrait-monomethod values are enclosed in solid lines. Heterotrait-heteromethod values are enclosed by broken lines.

behavioral observations). Thus, from the same sample of 500 third-grade school children, nine different measurements are obtained: three different self-reported psychological measures are administered, three assessments are obtained from their corresponding best friends, and three behavioral observations are obtained for each child by their teachers. The resulting correlation matrix of hypothetical relations involving the nine measures is presented in Table 4.2.

Four Critical Elements of the Matrix

In examining this set of results, Campbell and Fiske (1959) recommend that we consider three components of the matrix (we have added the fourth) before attempting to form an assessment of convergent and discriminant validity. These critical components are:

- 1 The *reliabilities* or *monotrait-monomethod values* for each of the three traits using the same measurement method, contained in the main diagonal of the matrix (parenthesized values), are considered first. Normally, a correlation of a variable with itself is 1.00. However, in displaying the MTMM, the monotrait-monomethod values are replaced with either the internal consistency reliability or test-retest reliability value of each measure. The reliabilities (either internal consistency or test-retest) of the measures must be strong enough to encourage further consideration of the data. If they are not, there is not much point to inquire about validity. As shown, the reliabilities of our hypothetical measures are reasonably strong.
- 2 Adjacent to the reliabilities in the main diagonal, the *heterotrait-monomethod* triangles are enclosed in solid lines. These entries reflect the correlation between two *different* traits (heterotrait) that are assessed by the *same* measurement method (monomethod). Thus, in the topmost data triangle of Table 4.2, we find the correlations between sociability and self-esteem, aggression and self-esteem, and aggression and sociability, when all three traits are measured by the self-rating method ($r = .50, .40$, and $.30$, respectively).
- 3 Next, we consider the *heterotrait-heteromethod* triangles, which are enclosed in broken lines. These values reflect the relation between *different* traits (heterotrait) assessed by *different*

methods (heteromethods) of measurement. In Table 4.2, the correlation between sociability as measured by self-rating (Method A) and self-esteem as measured by peer ratings is weak (i.e., $r = .11$).

- 4 The final entries to be considered are the *monotrait-heteromethod* values, which lie on the diagonal that separates the heterotrait-heteromethod triangles. These *validity diagonals*, as they are termed, reflect the association of presumably *identical* traits assessed by *different* methods of measurement. They are presented in bold type in Table 4.2.

Evaluating the MTMM Matrix

We use these various entries to assess construct validity. Practical tips should help in deciphering the mechanics of the matrix, although a definitive statistical analysis of the matrix acceptable to all researchers has yet to evolve. The MTMM is meant to drive home the idea that an observed test score consists of many parts, at a minimum, a “true score” (pure trait measure) and extraneous elements, including the specific method used to measure the trait. The idea of the MTMM originated during a period when personality trait theory dominated scale validation research (Campbell & Fiske, 1959). It is helpful to think of “traits” as what today’s researchers more broadly and intuitively refer to as constructs, or latent factors. “Method” does not refer to a research method or design in the strictest sense, but essentially a “modality of administration” or “method of measurement” (e.g., self- or peer ratings). Other measurement methods that could have been used in this example, given today’s technology, could include a Web-based survey, a cell phone interview, physiological and hormonal assessments, or brain scans. The same sample of participants would be required to complete all these methods to build a MTMM analysis. Although not limited exclusively to three trait factors assessed with three method factors, extending the matrix makes it increasingly unwieldy, as each new trait must be assessed by every method. For this reason, MTMM studies generally do not go beyond this number (Chang et al., 2012; Marsh, 1989).

In terms of classical test theory, the underlying theoretical traits represent true scores, and different assessment methods represent different sources of measurement error that influence the observed scores (Eid, 2000; Eid & Nussbeck, 2009; Hofling et al., 2009; Scullen, 1999). In practice, almost all research studies assess constructs using only a single measurement method (e.g., paper and pencil surveys, behavioral observations), and thus the contribution solely attributable to the method of measurement cannot be disentangled from trait variance in participants’ total scores. The MTMM is an important theoretical tool to help us visualize matrix of relevant relations, and to remind us of the many methodological artifacts that can affect the validity of obtained scores.

In their classic paper, Campbell and Fiske (1959) suggested the important requirements that should be met before we conclude our measures are valid indicators of a construct. The first is that the correlations in the validity diagonals (the monotrait-heteromethod values) be statistically and practically significant. This requirement is concerned with convergent validity. It is reasonable to require these values be strong because they are meant to express the association between different measures of the (presumably) identical trait. For most theoretical purposes, measurement method is considered theoretically incidental, and thus, should not affect the values of the traits or constructs of interest. We have convergent validity when there is a strong overlap among the various assessments of traits that are considered identical (and which presumably differ only in the way they were measured). Thus, in our example, all three of our measures of aggression should correlate strongly – this is a reasonable requirement if we want to infer that all three measures assess the same underlying construct. If they do not, it is possible that the

measurement method used to assess aggression, which plays no role in our theory, is impinging on our results. In Table 4.2, the correlational results satisfy the convergence requirement.

The second (critical) requirement of the MTMM is that the validity values exceed the entries in relevant heterotrait-monomethod triangles (which are bound by solid lines). This “common sense desideratum” (Campbell & Fiske, 1959, p. 83) requires the relations involving different measures of the same trait exceed the correlation between different traits that merely happen to share the same method of measurement. If this requirement is not met, we are left to conclude that systematic (but theoretically irrelevant) measurement error may be controlling outcomes to an unacceptable degree. As illustrated, the results of our hypothetical study are in accord with this second requirement. For example, consider the correlation of Aggression measured by Method A (self-ratings) and Method B (peer ratings) ($r = .58$). This monotrait-heteromethod (validity) entry exceeds some, but not all of the correlations involving different traits that happen to be measured by the same method; for example, the correlation between Sociability and Aggression measured by Method B (peer ratings) is higher ($r = .61$). As Campbell and Fiske observed, failing this requirement is “probably typical” of research on individual differences. Methods factors contribute substantially to obtained scores, so different traits assessed by the same method will often be significantly correlated.

A third requirement, which like the earlier one is concerned with discriminant validity, is that each validity value exceeds the correlations of the entries of the row and column in which it is located, within its respective heterotrait-heteromethod triangle. So, returning to the comparison of correlations using Methods A and B, we see that the first validity ($r = .70$) exceeds the values in its relevant row ($r = .20$ and $.18$) and column ($r = .11$ and $.10$).

In addition, using the language of the MTMM, the monotrait-heteromethod values should exceed associated heterotrait-heteromethod associations. This, too, is a reasonable requirement. It means that the relation between different measures of the (presumably) same trait should be stronger than relations between different traits assessed with different measures. Discriminant validity is called into question if the association of different traits determined by disparate measures exceeds that involving the theoretically identical trait that also is measured with dissimilar methods. In the example provided in Table 4.2, discriminant validity is supported in every case.

The final requisite of the MTMM technique is that the same patterns of trait relations are observed in the heterotrait triangles irrespective of measurement overlap; that is, the pattern of trait relations should be the same in the monomethod and the heteromethod blocks. Such a requirement would be met if method were truly incidental, the most desirable outcome. Trait-relation patterns should be maintained whether the traits are measured by the same method (as in the monomethod triangles) or by different methods (the heteromethod triangles).

Over the years, researchers have used these “rules of thumb” to interpret the outcome of MTMM research. Since its introduction, the technique has been used to study validities of many types of variables. For example, Bliwise et al. (2021) studied the utility of a wristwatch to measure sleep duration and continuity using the MTMM. Others have used the approach to facilitate scale development (Feeney-Kettler et al., 2019; Mearns et al., 2009), and others still used the matrix to investigate hoarding behavior (Meyer et al., 2013).

Even so, considerable controversy still surrounds the question of the proper statistical method to decompose the matrix, but these issues will not concern us here (for discussion of some of the issues involved in the analysis of the MTMM, see Crano, 2000; Kenny, 1995; Kiers et al., 1996; Schmitt & Stults, 1986). In a reprise of their technique more than 30 years after its publication, Fiske and Campbell (1992, p. 394) avoided the statistical slugfest surrounding the proper analysis technique for the MTMM, and instead observed that the “criteria proposed in the original article seems like a sound first step, especially when one has an extended research program and

sees the particular matrix as only a step toward constructing an improved set of measuring procedures.” This is sound advice. It reemphasizes the procedural nature of validity, the need for continual assessment and refinement of instruments, and the fact that validity ultimately is a data-informed judgment, not a state of being.

Threats to Measurement Validity

The different validation operations presented here should not be viewed as mutually exclusive. Each of these approaches provides valuable and necessary information and should be employed concurrently whenever possible. The aim of each is to supply information that enables construction of more and more refined measures of some psychological or behavioral attribute. It is through such operations that valid measures of theoretically relevant concepts ultimately emerge.

We hold that validity is not an all-or-none property, but rather a relativistic, descriptive quality that indicates the extent to which theoretically expected variations on the attribute of interest are reflected in the obtained measurements. This implies that the more valid the scale, the greater should be this correspondence. We acknowledge that the score a person receives on a measurement scale is never a pure, totally accurate picture completely determined by the attribute or construct in question, but rather is the product of a complex interaction of many factors, only one of which is indicative of the attribute or construct of theoretical interest. Methodologists have identified many factors that (more or less) systematically influence responses to questionnaires and other self-report measures. It is important that these threats to measurement validity (known variously as *response sets* or *response biases*) be understood and controlled, because they systematically influence participants’ responses, which may enhance reliability and validity artificially.

Response bias exists to the extent that the measurement operations themselves influence obtained results. The degree to which data are affected by these measurement artifacts, or by other construct-irrelevant factors, partially determines the scale’s validity. At the most extreme level, we could envision a situation of complete invalidity, in which the way questions are worded totally determines participants’ responses, independent of the content or meaning of the items themselves. Under ideal conditions, the opposite is sought; format and question wording are designed to have no independent influence, and only the content of the questions determines responses. This ideal is rarely met, but an understanding of some of the potential threats to attaining this ideal can enhance the degree to which it will be approximated.

Mood

Perhaps not surprisingly, respondent mood states may affect responses to a social query. An interesting example of mood effects is provided by Schwarz and Clore (1983), who asked people to tell them about the quality of their lives. The telephone survey was conducted over a wide geographic area, and some of the respondents were enjoying a bright sunny day when they answered. Others were in a part of the country where it was raining when the survey was done. Quality-of-life responses were different depending on the weather. The more pleasant the weather, the higher the reports of life quality. Obviously, the quality of one’s life may be affected somewhat by weather, but it is noteworthy to see how strongly such transitory effects can influence answers to questions that are meant to tap more stable states. In an interesting addition to this study, the researchers asked some of their respondents what the weather was like before beginning their questionnaire. Among these respondents, weather had no effect on

quality-of-life responses. Why? Presumably because when the source of their transitory mood state was brought to mind, they were able to discount it when answering the survey. This same result has been replicated in many different studies. When German respondents were asked to report on the quality of their lives, for example, they reported much higher quality-of-life scores if the German national soccer team had won their games in the World Cup Playoffs than if they had lost (Schwarz et al., 1987).

Social Desirability

In many measures of social variables, the respondent is asked to present a self-report concerning some more or less important self-revealing belief, value, attitude, or behavior. There are some situations in which an individual's actual beliefs, values, attitudes, or behaviors are not aligned with those approved by common social norms. Under such conditions, the respondent might be tempted to respond in a "socially desirable" way by misreporting and responding in a manner that is consistent with social mores. Variations in respondents' sensitivity to the demands of social desirability, or differences in people's perceptions of what is and is not socially desirable, can invalidate a scale. The findings of many early surveys of adolescent sexual behavior offer a good illustration of this point. In many of these reports, we discovered that a high proportion of adolescent men had engaged in sexual intercourse, whereas relatively few women of this age had done so. Such findings suggest several interesting possibilities, some of which seem quite implausible. It seems likely that a cultural value approving (or not disapproving) of sexually experienced males and virginal females was well learned by the time a person reached adolescence, and respondents' reports of their activities were at least in part a function of these learned values, rather than of actual behavior.

In attempting to solve the problems that a bias of this type can generate, it is useful to speculate on the factors that cause it. The social desirability response bias occurs because of a lack of self-knowledge on the part of the respondent (self-deception), his or her refusal to be completely frank or honest to others (impression management), or both. There is little that can be done when participants do not know themselves well enough to give answers based on fact and reality rather than some idealization formed in part by the demands of the society. On the other hand, administering scales and guaranteeing anonymity can combat problems attributable to a lack of candor. If respondents can be assured their responses will be anonymous and data confidentiality will be maintained, it seems more likely they would be willing to supply honest answers, even if these answers were contrary to established social norms or practices.

Language Difficulty

A more tractable problem arises when a verbal measure uses language that is different from that characteristically employed by the respondent sample. In Wilson and Patterson's (1968) scale of conservative attitudes, for example, the word *apartheid* appears. The definition of this word probably was quite apparent for Wilson's sample of British respondents. Personal experience in using this scale in the United States, however, showed a widespread understanding of this term should not be assumed. Because most scales of opinion are intended to be general instruments capable of being administered to a variety of respondents, it is advisable to determine in advance whether the meaning of the items that constitute the scale is the same as that originally intended. A preliminary interview of respondents drawn from the test population can supply an inexpensive, rapid, and usually accurate determination of the "understandability" of the items on a scale. If the language used is too difficult, or misinterpreted, an alternate wording should

be adopted. In brief, the language of the scale should be adapted to fit the language of the respondent group.

Extreme-Response Sets

There is some evidence in the attitude scaling literature suggesting that reliable differences exist among people in terms of their tendencies to employ (or to avoid) the extreme-response options of rating scales. Some people, it seems, characteristically select the middle parts of scales, whereas others characteristically employ the extremes. Researchers have correlated these response tendencies with intelligence, education, conservatism, dogmatism, and other personality attributes, but the results of such investigations are not clear-cut. There seems to be no way to design a scale that is impervious to this potential bias; however, we do have available statistical methods to determine the degree to which “extreme-response sets” occur on any given administration (see Nunnally, 1967, pp. 612–613; also, Cheung & Rensvold, 2000). Research generally suggests that the extreme-response tendency does not greatly affect validity, but it is conceivable that on some issues this bias could have an impact on results. Remember, however, a bias that skews the responses of only 5–10% of the sample could have a powerful effect on the outcome of a statistical analysis (Crano, 1997).

Acquiescence

People’s tendency to acquiesce to, or agree with, positively worded statements is the final stylistic response trait to be considered. A positively worded statement is one on which agreement indicates a position favorable to the attitude object under investigation. Research indicates that some people characteristically acquiesce or agree with positively worded statements. Known variously as the tendency to “guess true,” “acquiesce,” “agree,” or “yeasay,” this variable has generated more research on stylistic biases than any other. This interest was probably stimulated by the development of Adorno et al.’s (1950) theory of the “authoritarian personality.” The California F (for fascism) Scale was one of the principal measurement devices employed in the assessment of authoritarianism, and over time became popular, being used in hundreds of investigations. More than 50 years after the publication of the classic work of Adorno et al. (1950), research interest on authoritarianism and its measurement persists (Altemeyer, 1988, 1996; Duckitt & Bizumic, 2013).

All the items on the original F scale were worded positively, and thus, higher levels of agreement with the items resulted in high authoritarianism scores. Some researchers hypothesized that the characteristic to agree with positively worded statements, rather than the content of the statements per se, might be responsible for some respondents’ authoritarianism scores. Thus, the scale was not measuring authoritarianism so much as it was the tendency to agree with positively worded statements. To test such suspicions, specially constructed F scales consisting of positively worded items that were designed to be opposite in meaning to the original items were constructed. If item content were the major determinant of respondents’ scores, a strong negative correlation between the original and the derivative F scale items would be expected. Instead, *positive* correlations were obtained, thus supporting those who hypothesized the existence of an acquiescence response set.

Results of this type stimulated considerable research, which was focused not only on the F scale but also upon response sets in general (e.g., see Altemeyer, 1988; Block, 1965; Chapman & Campbell, 1957, 1959; Forsman, 1993; Jackson & Messick, 1962; van Herk et al., 2004). Much of this research employed item-reversal techniques of the type discussed above. A major

conceptual difficulty with this approach is that it is sometimes impossible to know when an item has truly been reversed. Consider the following item:

Capitalism represents the most equitable economic system.

One might reverse this item by substituting the word socialism or communism for capitalism; another possibility would change the word “equitable” to “inequitable.” Still other reversals of the original item are possible, and unfortunately, a person might logically agree or disagree with both the original and its reversal. In the present example, for instance, a respondent might feel that neither capitalism nor socialism nor communism was an equitable economic system. The predicted negative correlation between apparently reversed scales might not occur given enough reversals of this type. Considering the difficulties discussed, such a finding would not necessarily indicate the presence of an acquiescence response bias, but rather the failure to develop good reversals.

Interpretational problems of this type led Rorer (1965) to question the very existence of response sets. In an influential paper, Rorer argued that the response style question was a pseudo-issue and need not be considered a real danger to the validity of scales. This reaction to the complexities and interpretational difficulties of this research area is understandable, but probably too extreme. Campbell et al. (1967), following Rorer, were able to demonstrate convincingly the presence of response biases in both the F scale and selected subscales of the Minnesota Multiphasic Personality Inventory (MMPI). The effects of these stylistic biases were not great, but this is different from saying that they did not exist (Ferrando et al., 2004; Richardson, 2012).

Like all the other response set or stylistic bias issues, that of acquiescence has been shown to be a real phenomenon, but one whose potential effect on the validity of an instrument is not great; however, their impact on the outcome of an analysis could be sufficient to alter the conclusions we might draw from it (Weijters et al., 2010). If one is committed to the development of increasingly sensitive, reliable, and valid measurement instruments, even minor distortions should be eliminated.

Conclusion

To construct a measure that can be used to assess people’s attitudes, attributes, thought processes, or behavioral predispositions reliably and validly requires substantive knowledge, technical competence, and a certain intuitive feel for the ways that people think, feel, and act. Knowledge of correlational statistical techniques also is helpful, and with increasingly sophisticated analytic programs available at almost all research installations today, the computational work involved in these analyses is not technically daunting. Even more than in most areas of measurement research, “learning by doing” is crucial here. One learns how to write good items for an instrument by writing items and testing the quality of the resultant scales. One learns how to create good items from practice in developing and phrasing items. Coupled with strong theoretical knowledge of the phenomenon of interest, these experiences in measure construction (see Chapters 16 and 17) contribute to desirable psychometric properties of the new instrument.

References

- Adorno, T. W., Frenkel-Brunswik, E., Levinson, D. J., & Sanford, R. N. (1950). *The authoritarian personality*. New York: Harper.

- Altemeyer, B. (1988). *Enemies of freedom: Understanding right-wing authoritarianism*. San Francisco, CA: Jossey-Bass.
- Altemeyer, B. (1996). *The authoritarian specter*. Cambridge, MA: Harvard University Press.
- Ben-Porath, Y. S. (2012). *Interpreting the MMPI-2-RF*. Minneapolis: University of Minnesota Press.
- Bliwise, D. L., Chapple, C., Maislisch, L., Roitmann, E., & Burtea, T. (2021). A multitrait, multimethod matrix approach for a consumer-grade wrist-worn watch measuring sleep duration and continuity. *Sleep: Journal of Sleep and Sleep Disorders Research*, 44(1), 1–6.
- Block, J. (1965). *The challenge of response sets*. New York: Appleton-Century-Crofts.
- Bryant, F. B. (2000). Assessing the validity of measurement. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 99–146). Washington, DC: American Psychological Association.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Campbell, D. T., Siegman, C. R., & Rees, M. B. (1967). Direction of wording effects in the relationships between scales. *Psychological Bulletin*, 68, 293–303.
- Chang, L., Connelly, B. S., & Geeza, A. A. (2012). Separating method factors and higher order traits of the big five: A meta-analytic multitrait-multimethod approach. *Journal of Personality and Social Psychology*, 102, 408–426.
- Chapman, L. J., & Campbell, D. T. (1957b). Response set in the F scale. *Journal of Abnormal and Social Psychology*, 54, 129–132.
- Chapman, L. J., & Campbell, D. T. (1959b). The effect of acquiescence response-set upon relationships among F scale, ethnocentrism, and intelligence. *Sociometry*, 22, 153–161.
- Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology*, 31(2), 187–212.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago, IL: Rand-McNally.
- Crano, W. D. (1997). Vested interest, symbolic politics, and attitude-behavior consistency. *Journal of Personality and Social Psychology*, 72, 485–491.
- Crano, W. D. (2000). The multitrait-multimethod matrix as synopsis and recapitulation of Campbell's views on the proper conduct of social inquiry. In L. Bickman (Ed.), *Research design: Donald Campbell's legacy* (Chapter 3, pp. 37–61). Beverly Hills, CA: Sage.
- DeVellis, R. B. (2012). *Scale development: Theory and applications*. Thousand Oaks, CA: Sage.
- Duckitt, J., & Bizumic, B. (2013). Multidimensionality of right-wing authoritarian attitudes: Authoritarianism-conservatism-traditionalism. *Political Psychology*, 34(6), 841–862.
- Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika*, 65(2), 241–261.
- Eid, M., & Nussbeck, F. W. (2009). The multitrait-multimethod matrix at 50! *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 5, 71.
- Feeney-Kettler, K. A., Kettler, R. J., & Dembitzer, L. (2019). Application of a multitrait-multimethod matrix to social, emotional, and behavioral screening. *Journal of Applied School Psychology*, 35(1), 20–36.
- Ferrando, P. J., Condon, L., & Chico, E. (2004). The convergent validity of acquiescence: An empirical study relating balanced scales and separate acquiescence scales. *Personality and Individual Differences*, 37, 1331–1340.
- Fiske, D. W., & Campbell, D. T. (1992). Citations do not solve problems. *Psychological Bulletin*, 112, 393–395.
- Forsman, L. (1993). Giving extreme responses to items in self-esteem scales: Response set or personality trait? *European Journal of Psychological Assessment*, 9, 33–40.
- Gynther, M. D., & Buckhart, B. R., & Hovanitz, C. (1979). Do face-valid items have more predictive validity than subtle items? The case of the MMPI Pd Scale. *Journal of Consulting and Clinical Psychology*, 47, 295–300.
- Höfling, V., Schermelleh-Engel, K., & Moosbrugger, H. (2009). Analyzing multitrait-multimethod data: A comparison of three approaches. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 5, 99–111.
- Jackson, D. N., & Messick, S. (1962). Response styles on the MMPI: Comparison of clinical and normal samples. *Journal of Abnormal and Social Psychology*, 65, 285–299.

- Kenny, D. A. (1995). The multitrait-multimethod matrix: Design, analysis, and conceptual issues. In P. E. Shrout & S. T. Fiske (Eds.), *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske* (pp. 111–124). New York, NY: Psychology Press.
- Kiers, H. A. L., Takane, Y., & ten Berge, J. M. F. (1996). The analysis of multitrait-multimethod matrices via constrained components analysis. *Psychometrika*, 61(4), 601–628.
- Marsh, H. W. (1989). Confirmatory factor analyses of multitrait-multimethod data: Many problems and a few solutions. *Applied Psychological Measurement*, 13, 335–361.
- Mearns, J., Patchett, E., & Catanzaro, S. J. (2009). Multitrait-multimethod matrix validation of the negative mood regulation scale. *Journal of Research in Personality*, 43(5), 910–913.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.
- Meyer, J. F., Frost, R. O., Brown, T. A., Steketee, G., & Tolin, D. F. (2013). A multitrait-multimethod matrix investigation of hoarding. *Journal of Obsessive-Compulsive and Related Disorders*, 2(3), 273–280.
- Nunnally, J. C. (1967). *Psychometric theory*. New York: McGraw-Hill.
- Richardson, J. T. (2012). The role of response biases in the relationship between students' perceptions of their courses and their approaches to studying in higher education. *British Educational Research Journal*, 38, 399–418.
- Rorer, L. G. (1965). The great response style myth. *Psychological Bulletin*, 63, 129–156.
- Schmitt, N., & Stults, D. M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement*, 10, 1–22.
- Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Information and directive functions of affective states. *Journal of Personality and Social Psychology*, 45, 513–523.
- Schwarz, N., Strack, F., Kommer, D., & Wagner, D. (1987). Soccer, rooms, and the quality of your life: Mood effects on judgments of satisfaction with life in general and with specific domains. *European Journal of Social Psychology*, 17, 69–79.
- Scullen, S. E. (1999). Using confirmatory factor analysis of correlated uniqueness to estimate method variance in multitrait-multimethod matrices. *Organizational Research Methods*, 2, 275–292.
- van Herk, H., Poortinga, Y. H., & Verhallen, T. M. M. (2004). Response styles in rating scales: Evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology*, 35, 346–360.
- Weijters, B., Geuens, M., & Schillewaert, N. (2010). The individual consistency of acquiescence and extreme response style in self-report questionnaires. *Applied Psychological Measurement*, 34, 105–121.
- Wheeler, S. R., & Austin, J. (2000). The loss response list: A tool for measuring adolescent grief responses. *Death Studies*, 24, 21–34.
- Wilson, G. D., & Patterson, J. R. (1968). A new measure of conservatism. *British Journal of Social and Clinical Psychology*, 7, 264–269.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

PART II

Research Design Strategies

**Experiments, Quasi-Experiments, and
Nonexperiments**



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

5

DESIGNING EXPERIMENTS

Variations on the Basics

Chapter 2 introduced the concepts of internal and external validity (two important forms of validity to be considered in experiments) and provided some information on the basic structure of an experiment. Here, we expand on this basic structure to consider variations in the ways in which experiments can be designed, set up, and executed.

Basic Variations in Experimental Design

To summarize our earlier discussion, the classic “true experimental design” (Campbell & Stanley, 1963) involves the following steps:

- 1 Obtaining a sample of participants
- 2 Pretesting participants on the dependent variable of interest
- 3 Randomly assigning each participant to experimental or control groups¹
- 4 Carefully controlling for differences in the application of the experimental treatment between the two groups
- 5 Remeasuring the groups on the dependent variable at some time following the experimental manipulation

These steps are diagrammed in Figure 5.1. Variations on this basic structure include elimination of the pretest, the addition of multiple experimental treatments, and the repeated use of the same participants in all conditions of the experiment. In this chapter, we will discuss each of these modifications.

Pretest-Posttest Control Group Design

In the just described **pretest-posttest control group design**, both pretest and posttest are given to participants in control and experimental groups. Pretesting makes it possible to determine that participants assigned to different conditions of the study are initially equivalent in their response to the dependent variable at the outset. Ideally, after random assignment the two groups should be largely the same, on average, in their pretest scores within the limits of chance

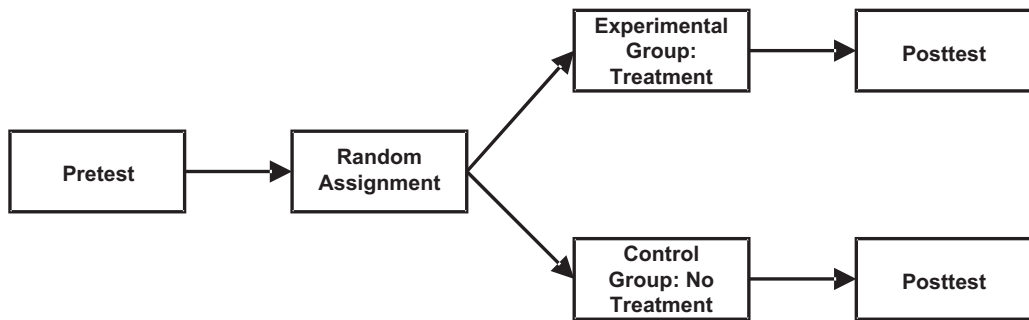


FIGURE 5.1 Pretest-posttest control group design.

variation. Random assignment also helps ensure that both conditions are similar on virtually all unmeasured extraneous variables. Pretests serve other purposes in controlling for individual differences, as will be discussed in this chapter and in quasi-experimental methods of Chapter 9.

Although pretesting has many benefits, this design is often altered to offset a potential problem that can be introduced by the pretest itself. This problem is termed **pretest sensitization**, or the possibility that the pretest can make participants in the experimental group unduly reactive to the treatment to which they are exposed. Under these circumstances, the effect of the treatment may be artificially enhanced – the treatment appears stronger, or more effective, than it would be when applied to people who were not exposed to the pretest. It also is possible that the pretest can dampen the effectiveness of a treatment if, for example, it alerts participants to the fact that the experimenter is trying to change their position on an issue and arouses their resistance to such change. Either pretesting effect is especially likely if the pretest is administered in the same session as the experiment itself, just prior to the introduction of the manipulation.

An example may help to clarify the sensitizing bias. Suppose that we administered a pretest of racial attitudes to a large sample of research participants. The treatment is a communication that we hope will influence participants to be less biased in their attitudes toward other racial groups. We administer the treatment and find it succeeded: Participants in the experimental group are much less biased on the posttest measure than are those in the control group (who did not receive the communication), though both groups were identical on pretest attitudes.

From these results we cannot be sure that the experimental communication alone would be an effective means of changing people's racial attitudes. It is possible the pretest might have sensitized the experimental participants to the treatment and thereby altered its effectiveness as reflected on the posttest. As a result of the pretesting, participants were unintentionally led to ponder the legitimacy of their racial beliefs. Those in the experimental group – already sensitized to the subject – were especially susceptible to the communication when they received it. Participants in the control group might also have been induced to think about their racial attitudes due to the pretest, but because they were not exposed to the communication treatment, they did not show as much attitude change at posttest assessment.

One way to reduce this problem would be to administer the pretest measure some time in advance of the treatment – at least a few days or weeks prior to participation in the experiment itself. Assuming that the participants do not make a strong connection between the earlier testing and the experiment, the passage of time may reduce any sensitization stemming from the pretest measure. However, this does not always work. If the pretest involves an important or emotionally loaded issue (such as in the racial attitudes example above, or a measure of self-esteem or self-concept), respondents may ruminate about their responses after they have taken

the test. Under these circumstances, it is possible that exposure to the pretest itself leads to a *change* in attitudes over time. Thus, when participants later enter the experiment, the pretest no longer reflects their current attitude or position. Even when this kind of sensitization is unlikely, it is not always possible to administer the pretest at a much earlier time. Participants may only be available for one measurement session.

Posttest-Only Control Group Design

To solve the problem of pretest sensitization, researchers usually eliminate the pretest as part of the experimental design. This is especially likely in experimental contexts in which a pretest would have to be administered in the same session as the experimental treatments and may bias results by sensitizing participants, thereby making them more (or less) susceptible to the treatment. The design solution to this problem is diagrammed in Figure 5.2. In the **posttest-only control group design**, only a posttest but no pretest is given to participants in control and experimental groups. Pretesting is used to determine that participants assigned to conditions were initially “equivalent” on the dependent variable, but random assignment allows the presumption of initial equivalence, by chance, of experimental and control groups. Thus, there is no need for pretesting to establish that the groups are the same prior to the introduction of the experimental treatment. Of course, this assumption rests on the use of a relatively large number of participants. Random assignment of a small sample of participants to the conditions of an experiment is not likely to produce initially equivalent groups on the variable of interest. As a general rule of thumb, we recommend that the sample of participants be large enough to randomly assign at least 25–30 individuals to each condition to have any hope of creating equivalent groups.

Solomon Four-Group Design

The posttest-only control group design has several advantages in reduced cost and effort, along with the avoidance of the pretest sensitization bias. However, it has one disadvantage in that it prevents the researcher from assessing the potential effect of pretesting on the experimental treatment under study. This disadvantage can be overcome by combining both the pretest-posttest and posttest-only control group designs. This is known as the **Solomon four-group design** (Solomon, 1949), used to determine treatment effect of experimental and control groups and the effect of pretest sensitization on the dependent variable. In this variation, participants are randomly assigned to one of four groups, two of which are pretested and two of which are

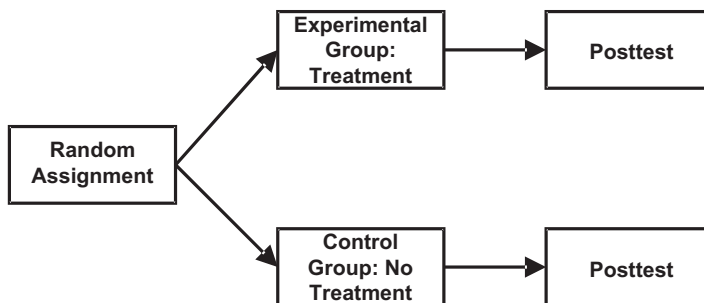


FIGURE 5.2 Posttest-only control group design.

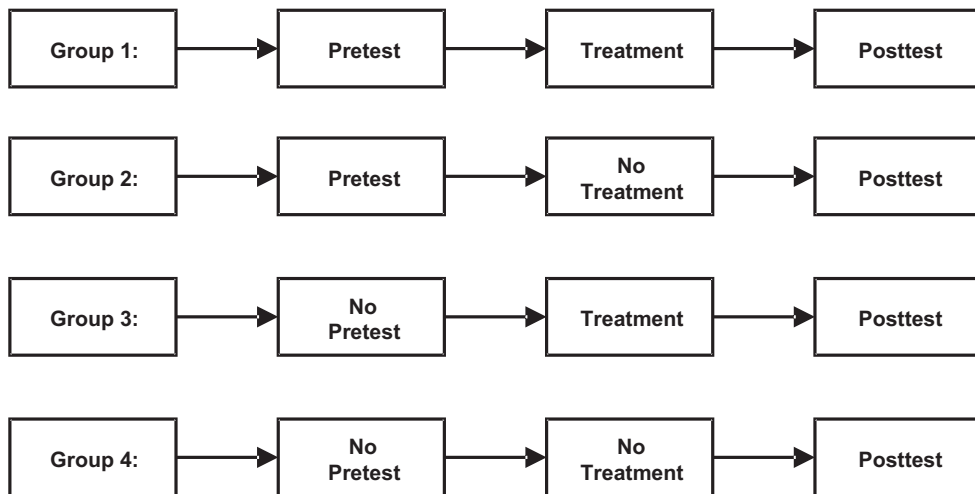


FIGURE 5.3 Solomon four-group design.

not. One of the pretested groups and one of the non-pretested groups are then exposed to the treatment; the other two groups are not, and this produces the division of participants represented in Figure 5.3.

The use of this design permits several effects to be evaluated in the same experiment. Equivalence of Groups 1 and 2 on the pretest suggests that the randomization procedure has been effective, and posttest results across all four groups can then be used to assess the effects of both the treatment variable and pretesting. Any effect of the experimental treatment, above and beyond testing and other rival factors, can be determined by comparing the posttest results of the two experimental groups (1 and 3) with those of the control groups (2 and 4). Effects of pretesting alone can be obtained by comparing the posttest scores of Groups 2 and 4, whereas any effects of pretesting on sensitivity to the experimental treatment can be detected in differences between Groups 1 and 3, who have received the same treatment but differ in their prior exposure to a pretest measure.²

A good example of the use of the Solomon four-group design is provided by Rau and colleagues (2010), who investigated the effectiveness of a sexual assault prevention program. The researchers were concerned that the preexisting attitudes of their participants (more than 1,000 Navy men at the completion of their basic training) might influence the results of the program, so they decided to measure these attitudes before the program. However, they also realized that the pretest might sensitize the men to the treatment. To assess both issues, they pretested approximately half the participants on their existing rape knowledge and empathy for rape victims before the program began; the remaining participants were not pretested but all participants were given these measures after the training. The prevention program enhanced the subjects' rape knowledge and empathy for rape victims, and reduced rape myth acceptance. These results held up even for the non-pretested subjects.

Expanding the Number of Experimental Treatments

Thus far we have been considering experimental designs in which the independent variable consists of the presence or absence of a single experimental treatment (the treatment-control

group design). However, there is no reason that more than one variation of an experimental treatment cannot be made within the same experiment so long as certain rules for enlarging the number of treatments are followed. First, instead of presence versus absence of the experimental treatment as the basic comparison, we can also perform comparisons with two types of treatment conditions that do not involve a control group. For instance, we might wish to know whether drug A is more effective than drug B in curing infections. This is a legitimate question, and a “no drug” control group need not be used.³ Furthermore, we may compare groups that have been exposed to different *amounts* or *kinds* of the independent variable. Then, in addition to determining whether the existence (presence or absence) of the treatment makes any difference in the posttest dependent variable, the researcher can determine whether variations in the treatment also make a difference. These variations may be quantitative – different amounts of the variable (e.g., 0 vs. 0.3 cc. vs. 1.0 cc. of a drug, or high vs. medium vs. low payment for an action) – or categorical – different ways of producing the independent variable (e.g., anxiety induced by social conditions vs. anxiety induced by physical danger). The number of variations of a given experimental treatment (including its absence as represented in a control group) is referred to as the number of *levels* of that variable. Thus, the basic treatment-control group design is a two-level experiment; introducing additional variations of the same experimental treatment (independent variable) expands the design to three, four, or more levels.

Factorial Designs

In addition to adding levels of an experimental treatment, an experiment can also be expanded to include more than one treatment factor. In a **factorial design**, levels of one independent variable or factor are combined with levels of other independent variable(s), to further expand the number of conditions in an experimental design. For complete interpretability of results, these variables should be conjoined so that the levels of each independent variable are combined with all levels of the other(s), to yield the various conditions.

Creating the design. With a factorial design, we can assess the effects of variation in one independent variable while systematically varying one or more other independent variables as well. Each independent variable manipulated in a factorial design is a “factor,” and the number of conditions (or “cells”) in the design is equal to the product of the number of levels of all its factors. For a two-factor design, for example, the combined treatments might be as shown in Table 5.1. In this illustration, Variable 1 has five levels, and Variable 2 has four levels. This would be designated a 5×4 factorial design. It consists of 20 (5×4) separate conditions, formed by the factorial combination of the two independent variables.

For a three-factor design, the two-factor case is repeated for each level of the third factor, as in Table 5.2. Theoretically, the number of variables that can be combined in one experiment

TABLE 5.1 Two independent variable (5×4) factorial design

		<i>Factor 2</i>			
		<i>Level 1</i>	<i>Level 2</i>	<i>Level 3</i>	<i>Level 4</i>
Factor 1	Level 1	Condition 1-1	Condition 1-2	Condition 1-3	Condition 1-4
	Level 2	Condition 2-1	Condition 2-2	Condition 2-3	Condition 2-4
	Level 3	Condition 3-1	Condition 3-2	Condition 3-3	Condition 3-4
	Level 4	Condition 4-1	Condition 4-2	Condition 4-3	Condition 4-4
	Level 5	Condition 5-1	Condition 5-2	Condition 5-3	Condition 5-4

TABLE 5.2 Three independent variable ($3 \times 2 \times 3$) factorial design

		Factor 3			
		Factor 2	Level 1	Level 2	Level 3
Factor 1	Level 1	Level 1	Condition 1-1-1	Condition 1-1-2	Condition 1-1-3
		Level 2	Condition 1-2-1	Condition 1-2-2	Condition 1-2-3
	Level 2	Level 1	Condition 2-1-1	Condition 2-1-2	Condition 2-1-3
		Level 2	Condition 2-2-1	Condition 2-2-2	Condition 2-2-3
	Level 3	Level 1	Condition 3-1-1	Condition 3-1-2	Condition 3-1-3
		Level 2	Condition 3-2-1	Condition 3-2-2	Condition 3-2-3

is limitless. Note, however, that the number of treatment groups, and therefore the number of participants required, increases multiplicatively as independent variables are added to the design. This fact automatically places some practical limits on the number of independent variables that can be included in any single study. Table 5.2, for example, presents a $3 \times 2 \times 3$ factorial design. Randomly assigning 25 participants to each cell of the design would require 450 participants, the product of $25 \times 3 \times 2 \times 3$. If we wished to add another factor involving three levels to the design and still maintain the same number (25) of participants per cell, we would need exactly three times as many participants as in the three-way ($3 \times 2 \times 3$) design of the table. Obviously, then, there are practical constraints on the number of conditions that can be run in a factorial design, so both the number of factors and the number of levels of each factor must be limited by available resources.

It is preferable, but not mandatory, that the same number of participants be used in each of the cells of the design, which represent the various combinations of levels of independent variables. Then, the effect of any single independent variable can be obtained by comparing dependent variable scores of all participants who were exposed to the first level of that independent variable with those exposed to the second level, the third level, and so on. A **main effect** is evidenced if overall mean differences are observed on a dependent measure as a function of one independent variable, while holding constant all levels of the other independent variable(s). Thus, equivalence of groups is maintained because an equal number of participants in all levels of the independent variable have been exposed to each of the levels of the other independent variables.⁴

Factor independence. Just as random assignment to experimental conditions presumes control over exposure to the independent variable, factorial designs presume that two or more experimental factors can be manipulated *independently*. Sometimes, however, the nature of two variables is such that the levels of one cannot be independent of the other. Suppose, for example, that one wanted to study the effects of high or low anxiety-producing conditions in combination with the physical attractiveness of an experimenter on the dependent measure, the complexity of children’s speech. Suppose further that our experiment required that anxiety be induced by the behavior of the experimenter. In such a design, it is unlikely that the two “independent” variables (anxiety induction and attractiveness) could be independently manipulated, because it is not reasonable to expect that participants would see the positive attractiveness of someone who was practically scaring them to death. This example indicates the importance of going beyond the mere mechanics of factorial combination. Some thought is required in combining independent variables to ensure that meaningful experimental conditions have been created. Although almost any set of independent variables can be combined on paper, the way they are perceived by our participants is critical.

Treatment interaction effects. There are several reasons why we might want to expand an experiment beyond a single independent variable. One of the most important of these is to allow for the identification of interaction effects, which are discussed later in this chapter. Conceptually, an **interaction effect** occurs when the effect of one experimental variable differs depending on the level of a second experimental variable. For example, suppose we are interested in manipulating both information about group consensus and the self-relevance of an issue to observe their combined effects on personal opinions and beliefs. We believe that high consensus will influence group members' opinions rather strongly, but only when the issue at hand is not highly self-relevant to the participant. When the issue is highly self-relevant, group members will resist the apparent consensus, and may even go so far as to adopt a more radical anti-group opinion than they held before the consensus became known. They might respond in this way because they resent the apparent restriction on their personal freedom that the consensus estimate implies (Burgoon et al., 2002). Low group consensus, on the other hand, is not expected to have much impact, no matter the self-relevance of the issue. We could not test this hypothesis by manipulating only one independent variable, so we combine two of them in a factorial design. The pattern of mean differences found in the dependent variable that occurs in the high self-relevance condition as a consequence of high or low consensus would be contrasted with those that occur when self-relevance is low.

Solomon four-group design: An example. We have already considered one 2×2 factorial design in our discussion of the use of pretesting in experiments. In the Solomon four-group design, two factors are systematically manipulated – presence or absence of a pretest, and presence or absence of the experimental treatment – each with two levels. Thus, the Solomon four-group design depicted in Figure 5.3 can be represented in factorial form as in Table 5.3. Using this design, the overall main effect of having received the experimental treatment versus not receiving it (ignoring pretesting conditions) is obtained by comparing posttest mean scores of participants in Row 1 (Groups 1 and 3) with those in Row 2 (Groups 2 and 4) of the table. The overall main effect of pretesting on posttest scores is obtained by comparing participants in Column 1 (Groups 1 and 2) with those in Column 2 (Groups 3 and 4). In obtaining the treatment effect, we can ignore the effect of the pretesting factor because both treatment and control groups have equal numbers of randomly assigned participants who have been pretested or not pretested. Thus, pretesting effects have been “held constant” across the two treatment conditions. The same logic holds when we wish to consider the main effect of the pretesting factor independent of the treatment factor. As can be seen, among the pretested group, an equal number of participants served in the treatment and control conditions, and this equality is found in the non-pretested group as well. Thus, treatment variations “cancel out” when assessing the main effects of pretesting. Finally, the pretest sensitization effect can be assessed by examining the *interaction* between the pretest factor and the treatment factor, as discussed in the following section.

TABLE 5.3 Solomon four-group design as a 2×2 factorial of treatment levels and pretest levels

		<i>Pretest main effect</i>	
		<i>Pretest</i>	<i>No pretest</i>
Treatment main effect	Treatment	Group 1	Group 3
	No treatment	Group 2	Group 4

Interaction Effects

We have shown how a factorial design allows us to examine the effect of each independent variable with the effects of all other independent variables controlled, or held constant. The overall effect of each independent variable is called the main effect of that factor (i.e., the effect of variations in the levels of that factor, when variations in the other factor or factors have been systematically controlled). However, the primary advantage of the use of factorial designs in which two or more experimental factors are combined lies in the opportunity to detect interaction effects, evidenced if the effect of an independent variable on a dependent variable is altered or moderated by variations in the level of other independent variable(s), while controlling for the main effects.

The combination of the pretest sensitization factor and treatment factor, discussed earlier in this chapter, is an example of a type of interaction effect. In the four-group design depicted in Table 5.3, sensitization might potentially produce differences in the effect of the treatment depending on whether the pretest had been present or not. In this case, the mean difference between Group 1 and Group 2 (pretest present) might be *greater* than the mean difference between Group 3 and Group 4 (no pretest). That would indicate that the pretest sensitized participants in such a way as to enhance the effect of the treatment manipulation. Thus, the effect of the treatment or no treatment is altered by the presence or absence of pretesting.

To return to an earlier example, suppose we wanted to test the effects of a sexual assault prevention program on Navy enlistees. We wanted to pretest all participants to ensure that those who received the program did not differ initially from those who did not. However, we were afraid that the pretest might sensitize the enlistees to the intervention. To test this possibility, along with the central question – the treatment’s efficacy – we factorially combine administration (or non-administration) of the pretest with either participation in the prevention program or some other (control) activity. The pretest and dependent measure consists of a scale of empathy for rape victims, whose score can range from 0 to 30.

We analyze these scores and find two statistically significant main effects: As shown in Table 5.4, we see that those who received the sexual assault prevention program were significantly more empathetic toward rape victims than those who did not. This main effect suggests that the treatment worked as we had hoped. However, in addition to the statistically significant treatment main effect, we also find a main effect of pretest. On inspecting the means, we see that those who received the pretest were significantly more empathetic than those who did not. However, looking carefully at the results in Table 5.4, we notice something interesting, namely that three of the four groups had identical mean scores. Empathy scores were enhanced by the treatment only if the participants had received the pretest. The main effect of the prevention was qualified by the level of another variable, in this case, the presence or absence of the pretest. This “qualification” of one variable’s effect by another is called an interaction effect, which supplies considerably more valid insights into the efficacy of the prevention treatment program than either of the two independent variables considered in isolation.

The interaction effect obtained in our Solomon four-group example is important for both theoretical and methodological reasons. For methodological reasons, it was important to know that the pretest apparently sensitized participants to the treatment. For theoretical reasons, it was important to learn that the treatment had no effect when administered without a pretest measure that sensitized participants to the treatment.

The Solomon design is not used often in social science research, in part because the issue of pretest sensitization often does not arise. However, understanding how levels of one manipulated variable affect participants’ reactions to different levels of other independent variables is a regular feature of experimental research. Predicting and understanding interactions between

TABLE 5.4 Mean “empathy for rape victim” scores as a function of pretest administration and participation in the sexual assault prevention program

		<i>Pretest main effect</i>		
		<i>Pretest</i>	<i>No pretest</i>	
Treatment main effect	Prevention treatment	20	5	12.5
	No treatment	5	5	5
		12.5	5	—

and among independent variables on dependent variables is crucial to enlarging our understanding. Almost always, study of the interactions among independent variables of theoretical interest enhances our understanding of the phenomena we are investigating. For example, suppose we were interested in the effects of various styles of leadership on the productivity of small work teams and, further, theorized that the stress of outside competition interacted with these leadership effects. To address this issue, we might design an experiment in which four-person teams are assembled; given tools, materials, and instructions; and tasked with producing widgets.

The groups could be set up so that each team’s leader was trained to act in either a democratic or an authoritarian manner. Leadership style thus becomes an experimental treatment, or factor, with two levels. The second factor, which we label competitive stress, has three levels. This factor is created by informing participants that they are competing with an outside company, whose production figures are made available to them. The sales information the teams receive is the second independent variable. It indicates that the “other” company always seemed much more productive, about the same as, or much less productive than, the participants’ own work team. To enliven the proceedings, we could promise a \$500 prize to the most productive group. Notice that all participants are promised this reward, so it is a constant across teams, and thus, it is not an experimental treatment (or independent variable).

By combining the two independent variables factorially, the experiment takes the form of a 2 (leadership: democratic vs. authoritarian) $\times$ 3 (competitive stress: low vs. medium vs. high) factorial design. In this setup, instead of using individual participants, the productivity of each four-person work team is the unit of analysis. The number of widgets produced by each work team in each of the six conditions formed by the factorial combination of the two independent variables is the study’s index of group productivity, the dependent measure. Suppose that we randomly assigned 15 teams (of four workers each) to each of the six conditions formed by the factorial combination of our two experimental factors. The average team productivity (number of widgets produced) in each condition is depicted in Table 5.5. A quick scan of the results indicates that neither the Leadership nor the Stress factors alone produced a substantial main effect: Across the two levels of leadership style, the groups exposed to High competitive stress produced 35 pieces of widgets on average, just as did each of the groups exposed to Medium and Low stress. Similarly, the democratically run groups produced 35 widgets on average, neither more nor less than the authoritarian groups.

Although there is no evidence of a main effect for either independent variable or the dependent variable, by simply looking at the data pattern (always a good idea) it becomes clear that the manipulations did influence widget productivity. Leadership style per se had no differential impact on productivity, but *in combination with* levels of competitive stress, its influence on productivity was qualified. The interaction of these two independent variables on the dependent variable suggests that under the condition of low stress, a democratically structured group will

TABLE 5.5 Mean number of widgets produced as a function of leadership style and competitive stress:
Example of a crossover interaction

		Competitive stress			
		Low	Medium	High	
Leadership style	Authoritarian	25	35	45	35
	Democratic	45	35	25	35
		35	35	35	–

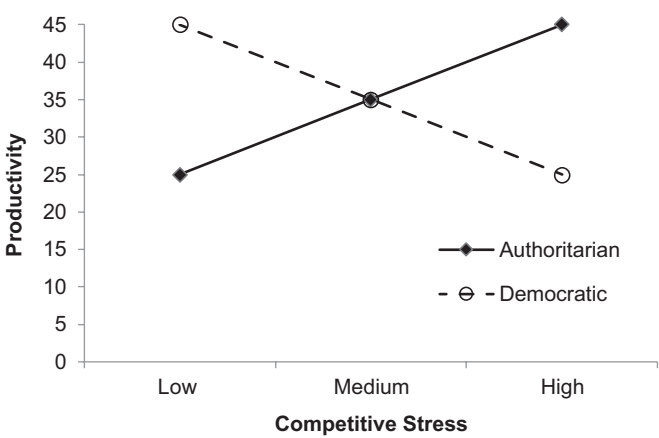


FIGURE 5.4 Effects of competitive stress and leadership style on productivity (see Table 5.5).

out-produce groups led in a more authoritarian fashion. However, as the stress of competition increases, the authoritarian teams increased productivity, whereas the democratic teams became less productive. Because of this interaction effect, we cannot specify the nature of the effect of leadership style on production without knowing the competitive stress conditions under which the work team is operating.

Common forms of interaction effects. Interactions can take many different forms. The most common interaction forms are *divergent* interactions and *crossover* interactions. The productivity data of Table 5.5 provide an illustration of a crossover form of interaction. The basis for this terminology is evident from the graphic portrayal of the results in Figure 5.4. As shown in this figure, the productivity levels of the authoritarian and democratic groups cross over, in this case at the medium level of competitive stress.

A divergent interaction effect is illustrated in the data of Table 5.6, which is depicted graphically in Figure 5.5. Note that the divergent effect could just as well take on the appearance of a *convergent* interaction if the levels of competitive stress were reversed – i.e., if the High level were presented first and the Low level last on the horizontal axis. Convergent and divergent interactions are simply alternative forms of the same interaction effect.

As shown in this illustration, democratic and authoritarian groups are equally productive under low levels of stress, but as stress increases, the production rates of the groups diverge: the democratic groups become less productive, the authoritarian groups more so. A main effect of leadership style also becomes evident in this example – across stress conditions, the authoritarian groups produced more widgets overall than did the democratic groups. However, interpretation

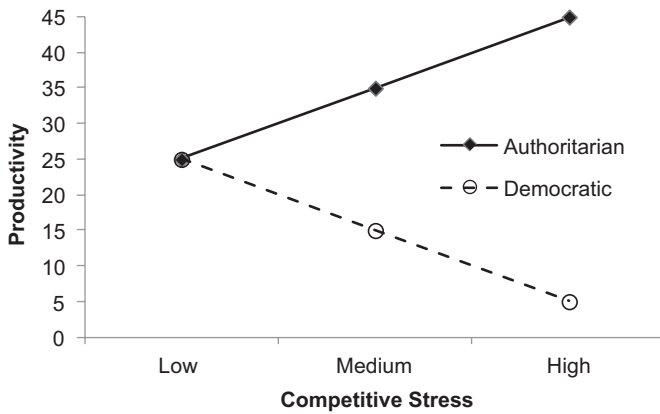


FIGURE 5.5 Effects of competitive stress and leadership style on productivity (see Table 5.6).

TABLE 5.6 Mean number of widgets produced as a function of leadership style and competitive stress:
Example of a divergent interaction

		Competitive stress			
		Low	Medium	High	
Leadership style	Authoritarian	25	35	45	35
	Democratic	25	15	5	15
		25	25	25	—

of this main effect must be tempered in light of the interaction effect obtained. Authoritarian groups did not out-produce democratic groups under all conditions, only when competitive stress was medium or high.

Both forms of interaction effect should be contrasted with what happens when no interaction is present. To stay with our example, it may be that authoritarian leaders outperform democratic leaders regardless of stress condition, and performance also decreases as stress goes from low to medium to high. In this case, we have only two main effects and these effects are *additive*. The resulting graphical representation is two parallel lines, as illustrated in Figure 5.6.

Moderator variables as interaction effects. In our discussion of causal relations in Chapter 2, we discussed the role of moderator variables. Recall that a moderator is a factor that alters (inhibits or enhances) the effects of a particular independent variable on a dependent variable. That is, the causal relation between A and B is modified by the presence or absence of a third variable, C. When both A and C are included in an experimental design as factorially combined independent variables, we can test whether A and C interact in determining the outcome on the dependent measure. The presence of an interaction is essentially a test of the moderator variable hypothesis. Thus, in our productivity example, we can say that the causal effect of authoritarian versus democratic leadership style was *moderated by* competitive stress. That is, the level of stress influences whether or not leadership style has a positive effect on group productivity.

In this example, the moderator variable is an environmental condition that can be manipulated by the experimenter and is subject to random assignment. In many cases, however, a hypothesized moderator variable is a trait or characteristic of individuals that cannot be manipulated or controlled by an experimenter. For example, the effectiveness of democratic leadership

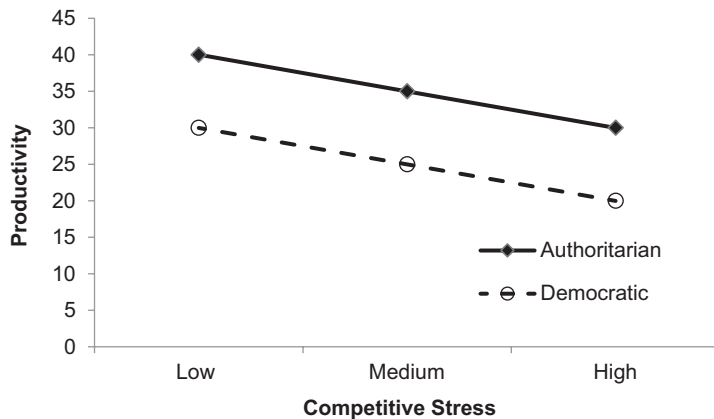


FIGURE 5.6 Effects of competitive stress and leadership style on productivity.

may be influenced by group members' high or low need for structure. Those who are low in need for structure may perform better under a democratic leader than under authoritarian leadership, whereas those high in need for structure may not do as well under democratic leadership. Because need for structure is a personal characteristic that is not under experimenter control, it cannot be manipulated as an independent variable. But the possibility of an interaction between leadership style and need for structure can still be tested by measuring participants' need for structure and then using level of need as a *blocking variable*, as discussed in the next section.

Block Designs: Incorporating a Nonexperimental Factor

As was mentioned, the effectiveness of random assignment in creating initial equivalence among our various experimental conditions depends upon the “law of large numbers.” Only when a large pool of participants is available can random placement guarantee a cancelling out of uncontrolled sources of variations across cells of the experimental design. If the initial pool of participants is relatively small and diverse, the researcher may be concerned that randomization may not distribute participants evenly among groups. When only a small initial pool is available, and particularly when this pool includes a few extreme cases, random assignment may not produce the desired equivalence between conditions.

Further, a researcher might be interested in knowing whether different types of individuals respond differently on the dependent measure, or whether such factors moderate a treatment effect. Generally, researchers are unable to manipulate individual difference variables such as participant intelligence, sex, religion, or country of birth, so these are necessarily *nonexperimental variables*. However, such variables can be combined with experimentally manipulated variables in factorial designs. In most cases, nonexperimental factors are characteristics of the participants themselves (e.g., gender, personality traits, intelligence), but they could also be environmental conditions (such as time of day). When one or more variables in a factorial design are not experimentally controlled, this is known as a **block design**.⁵

Blocked Randomization

When treated as a factor in a factorial design, a nonexperimental variable must be categorical. Ideally, participants are measured and sorted into categories *before* assignment to experimental

groups. This variation in the procedure of random assignment is called a **blocked randomization** design.

In this design, the members of the participant pool are first classified by levels on the non-experimental variable (for example, dividing participants into groups such as high, medium, or low on some measure of intelligence, or sorting participants by sex). Then, from *within* each of these groupings or “blocks,” participants are randomly assigned to experimental treatment groups so that each group is intended to have an equal number of members for each level of the blocked variable. As long as this assignment of pre-classified participants is determined by chance, the assumption of random assignment is not violated, and the initial equivalence of experimental groups is assured.

One form of a blocked design makes use of a pretest on the dependent variable (usually a pretest that has been administered some time prior to the experimental session). All participants in the pool are first categorized into groups based on pretest scores. Then participants within each of the groups, or blocks, are randomly assigned to experimental and control conditions, as diagrammed in Figure 5.7.

A blocked design is a form of factorial design. The blocking factor is crossed with the experimental factor (or factors, if there is more than one independent variable) in a mixed factorial design. Experimental factors are those that can be manipulated by the researcher and randomly assigned to participants. Blocking variables are not manipulated or randomly assigned – they are characteristics with which the participant comes to the experiment. Blocking increases experimental control by assuring that differences among participants on the blocked variable are approximately equalized across experimental conditions. In terms of interpretation of main effects, causal inferences could be made regarding the effect of the experimental factor, but not the blocked factor, on the dependent variable. Systematic differences related to the blocked factor are said to be associated with the dependent measure, but not necessarily in a causal fashion, because the blocked factor was not manipulated independently by the researcher.

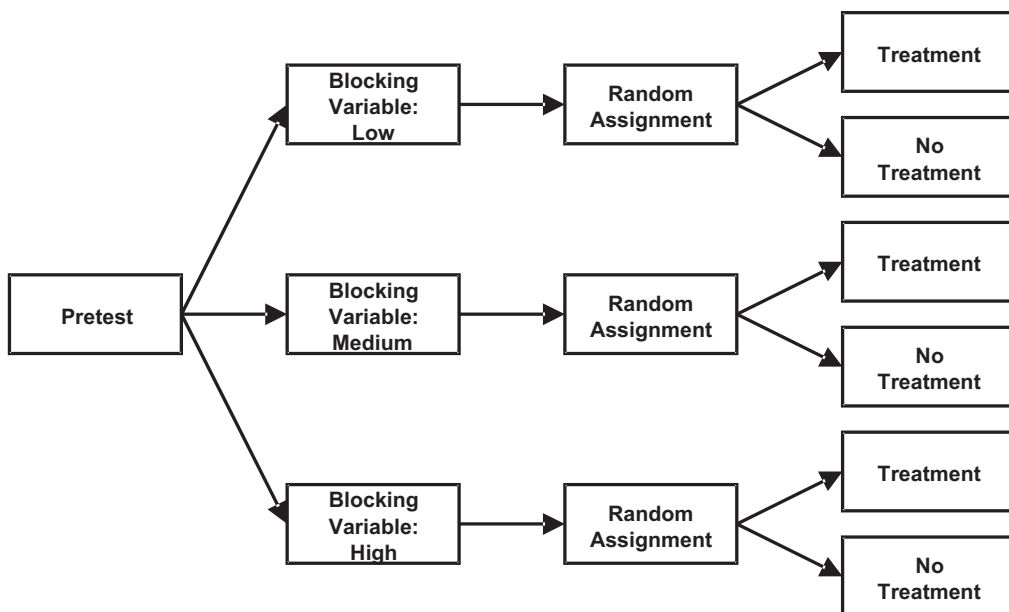


FIGURE 5.7 3×2 design of blocked variable and experimental variable.

Although blocking participants on the nonexperimental factor first and then randomly assigning within blocks is the ideal way to create a block design, it is often the case that researchers decide to take certain nonexperimental variables into account only after participants have already been randomly assigned to experimental treatment conditions. In this case, the process of creating a blocked factorial design is reversed – participants within each experimental condition are then categorized (blocked) according to their standing on the nonexperimental variable. This is generally considered acceptable as a way to incorporate a nonexperimental variable into a factorial analysis, but it is less than ideal. The advantage of blocking first and then randomizing assignments within blocks is that it guarantees that the levels of the nonexperimental variable will be evenly (or proportionally) distributed across the experimental conditions (e.g., every experimental condition will have an equal number of high, medium, and low scoring participants, or every condition will have 2/3 females and 1/3 males, etc.) When classifying participants is done after the fact, such equal distribution is not guaranteed. More often than not (unless a very large number of participants is involved), experimental conditions will differ in the proportion of participants at each level of the nonexperimental factor, creating factorial combinations with unequal *ns* (and inadvertent correlation between our nonexperimental and experimental variables). Although this can be handled statistically, it weakens the effectiveness of a mixed design to some degree.

Increasing statistical power. One potential reason to include nonexperimental variables as factors in a design is to enhance the power of our design to detect meaningful differences as a function of the experimental variable(s). The variance of a nonexperimental factor on our dependent variable is controlled, or accounted for, when that factor is a part of the experimental design. If the factor is not a part of the design (i.e., not systematically distributed by the researcher), its variation is unexplained and must be classed as error or “unaccounted for” variance in the dependent variable. This unaccounted error variance lowers the power of our design to detect differences between groups, and hence may lead to Type II error, mistakenly failing to reject the null hypothesis (Chapter 2).

Detecting interactions. Mixed designs also permit testing interactions between the experimental variable(s) and participant characteristics. For instance, it may be that the effectiveness of a particular persuasive communication depends on the level of intelligence of those who hear it. A complex argument may be processed, understood, and accepted by recipients who are high in intelligence but misunderstood or rejected by those with lower levels of intellectual capacity or cognitive skills (see McGuire, 1997; Rhodes & Wood, 1992). A simplistic argument, on the other hand, may appeal to people of average intelligence but be dismissed by high IQ individuals. If intelligence of the audience is ignored, it may appear that across all participants, no difference is found in the effectiveness of complex or simple arguments. However, if participant intelligence is measured before the experimental communication is delivered, and participants are then divided into high, medium, and low IQ blocks prior to random assignment, differences in effectiveness of the two types of communication could be observed between the different IQ groups. Using a blocked variable may enable us to learn more about the intelligence conditions under which different types of communication are effective. In this example, intelligence can be recognized as a *moderator variable*, as discussed earlier. An interaction between IQ level and message complexity means that the effect of communication complexity on attitude change is altered by whether the recipient is high or low in intelligence.

It also is possible to conduct a fully blocked design using a variant of the experimental designs discussed in this chapter. Due to ethical and practical restrictions on the capacity to manipulate some types of variables and attributes (e.g., participant demographic characteristics), the research might involve only blocked variables, which technically reduces it to a nonexperimental design.

Such a study might involve, for example, a 2×2 factorial combination of participants' gender and intelligence level as factors affecting performance on a verbal ability test. Neither of these factors are "independent" of co-occurring extraneous variables that may be responsible for confounding the observed differences in verbal ability between gender groups. For example, if it were the case that women tend to be better listeners or more attentive in speech acquisition, these unmeasured confounds, rather than gender, may be responsible for higher verbal ability in women over men. Although blocked variables may be used as factors in an experimental design, such designs do not have the internal validity of a "true" experiment because participants are not randomly assigned to levels of the factors. Thus, the ability to make cause-and-effect conclusions about the relations involving the design factors and the dependent variables is limited.

Repeated-Measures Designs and Counterbalancing

In most experimental designs, each participant is exposed to one and only one condition. Although a condition may represent the factorial combination of a number of experimental treatments, it is still true that each participant is assigned to only one cell of the experimental design. In these cases, each independent variable is known as a *between-subjects factor*. However, the same participants may be exposed to more than one level of an independent variable or combination of treatment conditions. This is known as a **repeated-measures design** because the same participants are repeatedly measured on the dependent variable, usually after each treatment exposure at each level of the independent variable(s). An independent variable manipulated in this way is called a *within-subjects factor*.

The pretest-posttest control group design is an example of a repeated-measures design. In this design, all participants are measured on the dependent variable once prior to the treatment and then again after they have been exposed to the independent (treatment) variable. Thus, the pretest can be used as a no-treatment base of comparison for assessing the treatment effect as reflected in the posttest. In other types of repeated-measures experiments, all participants are exposed to a series of different treatment conditions, one after another, and the dependent variable is administered after each exposure. Every participant may be exposed to all possible treatments in a predetermined order, or better yet, to minimize carry-over effects of any single order, they may be randomly assigned to different sequences of treatment orderings.

Just as pretest sensitization effects can influence the results of a pretest-posttest experiment, the effects of one level of a treatment may carry over in some way to influence the effects of succeeding treatments. Thus, the order in which treatments in a repeated-measures study are administered may have a major effect on its results. There is no way that order effects can be eliminated from the pretest-posttest design because, by definition, the pretest measure has to be taken first, before any other experimental treatment is administered. With successive treatments, however, varying the order in which different participants receive the various treatments can control for the impact of treatment order effects. For maximum interpretability of results not confounded by a particular treatment sequence, the same set of treatments should be presented to different participants in *counterbalanced* order.

In **complete counterbalancing**, to control for treatment order effects, sequences involving every combination of treatment orderings are determined, with each participant randomly assigned to one of the sequences. If a study involves only two conditions, complete counterbalancing is manageable. Half the sample could be assigned to receive the first treatment, followed by the second treatment. The remaining half of the sample would receive the treatments in reversed order. Extending to three treatment conditions, complete counterbalancing can be achieved with six different orderings, as shown in Table 5.7.

TABLE 5.7 Complete counterbalancing with three treatment conditions

		<i>First</i>	<i>Second</i>	<i>Third</i>
Sequence	1	A	B	C
	2	A	C	B
	3	B	A	C
	4	B	C	A
	5	C	A	B
	6	C	B	A

TABLE 5.8 A partially counterbalanced design

		<i>Order of communication treatment</i>			
		<i>First</i>	<i>Second</i>	<i>Third</i>	<i>Fourth</i>
Sequence	1	A	B	C	D
	2	B	D	A	C
	3	C	A	D	B
	4	D	C	B	A

Repeated-measures designs may entail more than three treatment levels from an independent variable but counterbalancing then becomes more complex and rather unwieldy. To control for order effects in such cases, **partial counterbalancing** is a compromise to complete counterbalancing that uses a *Latin square* design, to take into account the fact that it may be cumbersome to assign participants to every possible sequence of treatment orderings.

For instance, if each participant is to be exposed to four different messages – one from a pleasant communicator (treatment A), one from a neutral source (B), one from an unpleasant communicator (C), and one for whom no source is identified (D) – a counterbalanced ordering could be achieved by varying the order (or sequence of presentation) as in Table 5.8. In this scenario, after exposure to each treatment message, participants are immediately measured on the dependent variable four times. The dependent variable is used to determine if there are changes in mean scores from treatment to treatment.

Notice that partial counterbalancing does not involve using every possible order of all conditions (for designs with four or more treatments, this would run into a nearly unmanageable number of combinations). The requirements of partial counterbalancing are met if: (1) each treatment occurs once and only once in each position, and (2) each treatment is immediately preceded by every other treatment once and only once across the presentation orderings. In our four-condition example, treatment A is the first treatment (in Sequence 1), preceded once by B (Sequence 4), once by C (Sequence 3), and once by D (Sequence 2). The only other requirement for partial counterbalancing is that the number of participants randomly assigned to each sequence be equal, so that each ordering is used an equal number of times. Using this procedure, each sequence of participants receives one of the different orderings, and sequence becomes a type of blocking variable in the overall experimental design. When counterbalancing is used in a repeated-measures design, researchers are interested in the mean differences obtained among the four treatments, after collapsing results across the different sequences (not whether the four sequences yield different mean scores).

Repeated-measures designs, with counterbalancing, have the advantage of assuring the equivalence of groups exposed to different treatments because each participant, in effect, serves as his or her own “control group.” Because participants are exposed to all possible conditions, participant differences in characteristics and attributes across conditions are held constant, and if counterbalancing is successful, any observed differences across conditions should be attributable to receipt of the different treatments. However, repeated use of the same participants is just not possible for all types of experimental variables. In many cases, one level or combination of variables would severely interfere with, or preclude, administering any other treatment combinations.

For some variables, effects are cumulative. In studies involving intake of different amounts of drugs or alcohol, for instance, the experimenter could not administer a second dosage level to the same participant until all traces of the first dose were eliminated from that participant’s system; otherwise, the repeated doses would accumulate and obliterate any differences in the effects of dose level. Other experimental treatments are such that once participants have been exposed to one of them, their fatigue, level of awareness of experimental purposes or procedures, or practice effects may make them unsuited for use in further experimental conditions. However, there are some types of experimental manipulations that lend themselves to repeated-measures designs more readily. For instance, people in real life often are exposed to information about several different persons or to multiple news stories in succession on a particular topic. So, in experiments where the independent variable involves different content of information about persons or events, exposing the same participant to different levels of the manipulation may be reasonable. Even in these cases, however, the order in which a particular condition is received may make a difference, but counterbalanced design makes it possible to assess such order effects and take them into account in interpreting the effects of the experimental treatment(s).

Conclusion

In this chapter we started with the basic two-group pretest–posttest experimental design and demonstrated how elements can be added to or subtracted from that basic design to create useful variations in the structure of an experiment. The levels of an independent variable can be expanded from two conditions to multiple variations. More than one independent variable can be manipulated in an experiment using factorial designs. Pretesting can be excluded or included and used either as a blocking variable or in a repeated-measures design. Deciding among these different design features must be done based on the purposes of the experiment (including theories about potential interaction effects among different variables), and one’s knowledge or intuitions about factors such as sensitization, carry-over, and order effects.

Whatever form the experimental design takes, the design stage is still just the blueprint of the experiment itself. The design tells us what variables are being manipulated, at how many different levels, and in what combinations. Once this blueprint has been put into place, the real task is constructing the conditions and procedures through which the intended design will be implemented – i.e., the *operations* of the experiment. As with any construction, building each experiment always involves some unique features, decisions to be made, and problems to solve. However, there are some general principles and guidelines for constructing experimental procedures that can help maximize the internal validity of the results of the experiment and contribute to external validity as well. The following chapter covers some of these general principles in a step-by-step approach to conducting a laboratory experiment.

Notes

- 1 In much of our discussion, we assume that experiments are conducted with individual persons as the unit of analysis. However, in some cases, experimental treatments are delivered not to individuals independently assigned to groups, but to groups of persons (e.g., small work groups or even whole classrooms) randomly assigned to one or another condition of the experiment. In this case, the group, rather than the component individuals within the group, becomes the unit of analysis. More discussion of the issue of randomizing individuals or groups will be provided in Chapters 6 and 19.
- 2 A review of research (Lana, 1959) making use of this four-group design indicated that loss of generalizability of findings because of pretesting effects was minimal. However, Rosnow and Suls (1970) found that the pretesting effect operates differentially in groups of volunteer and nonvolunteer subjects. Because this issue remains unresolved, investigators are urged to exercise caution when interpreting studies that make use of pretests, especially in contexts in which pretest sensitization is a reasonable possibility.
- 3 However, by not having a control group, it is impossible to determine whether either drug is more effective than nothing at all.
- 4 Strictly speaking, we can employ different numbers of subjects in the various treatment combinations of a factorial design, and at times this may be desirable. If certain assumptions of proportionality are met, common statistical techniques (e.g., Winer, 1971) will allow us to estimate treatment effects. However, both analysis and interpretation are simplified when an equal number of participants are employed in each condition.
- 5 Some textbooks refer to this inclusion of nonexperimental factors as “quasi-experiments,” but this is a misnomer because the researcher has no influence on the participant characteristics. See Chapter 9 for discussion of quasi-experiments.

References

- Burgoon, M., Alvaro, E. M., Grandpre, J., & Voloudakis, M. (2002). Revisiting the theory of psychological reactance: Communicating threats to attitudinal freedom. In J. Dillard & M. Pfau (Eds.), *Handbook of persuasion* (pp. 213–232). Beverly Hills, CA: Sage.
- Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Boston, MA: Houghton Mifflin.
- Lana, R. E. (1959). Pretest-treatment interaction effects in attitude studies. *Psychological Bulletin*, 56, 293–300.
- McGuire, W. J. (1997). Creative hypothesis generating in psychology: Some useful heuristics. *Annual Review of Psychology*, 48, 1–30.
- Rau, T. J., Merrill, L. L., McWhorter, S. K., Stander, V. A., Thomsen, C. J., Dyslin, C. W., ... & Milner, J. S. (2010). Evaluation of a sexual assault education/prevention program for male U.S. Navy personnel. *Military Medicine*, 175, 429–434.
- Rhodes, N., & Wood, W. (1992). Self-esteem and intelligence affect influenceability: The mediating role of message reception. *Psychological Bulletin*, 111, 156–171.
- Rosnow, R. L., & Suls, J. M. (1970). Reactive effects of pretesting in attitude research. *Journal of Personality and Social Psychology*, 15, 338–343.
- Solomon, R. L. (1949). An extension of control group design. *Psychological Bulletin*, 46, 137–150.
- Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.

6

CONSTRUCTING LABORATORY EXPERIMENTS

In the preceding chapter, we discussed the design of experimental studies at a relatively abstract level to introduce basic principles of planning and constructing experiments. In the present chapter, we will become somewhat more concrete, and consider how to *implement* an experimental design in terms of the basic construction of a laboratory experiment, and the different forms of experimental manipulations that could be pursued. Along the way, we also reflect on other aspects of the experiment that, while not formal features of the design, can nevertheless have a powerful impact on the interpretation of a study. Although the chapter focuses on constructing laboratory experiments, some of the concepts also are applicable to conducting experimental research outside the laboratory.

Steps for Constructing an Experiment

Figure 6.1 presents a flowchart for designing and constructing a typical laboratory experiment. In developing a laboratory experiment, the researcher attempts to create an “alternate universe” – a small but self-contained environment in which participants are exposed to an experimental manipulation and is the location where the main “action” of the study takes place. Each step in the construction of the experiment has to be planned and carefully controlled by the experimenter.

Identify Participant Pool

The first step in developing a study is arranging for the availability of a pool of eligible participants – the sample of people who you plan to later recruit for the study. This first step will dictate many of the later decisions the experimenter must make – the form of experimental manipulation, the types of outcome measures to be administered, the extent that the researcher will play a role in the experimental context, and so on. We will discuss some of the ways in which participants may be recruited for participation in experiments later in this chapter. At this point, let us assume that a pool of participants is potentially available for the study. It is the researcher’s role to define the participants eligible to take part in the investigation. For some purposes, owing to practical or theoretical reasons, researchers may wish to recruit individuals with

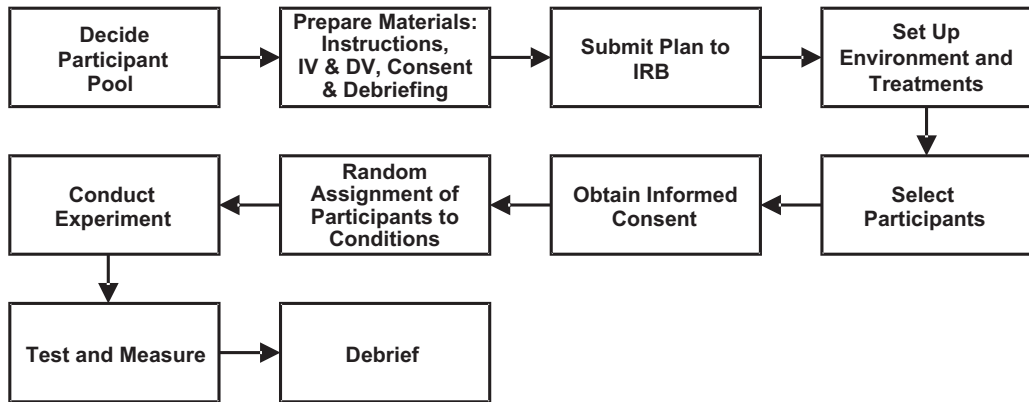


FIGURE 6.1 Framework for constructing an experiment.

particular demographic characteristics, such as only male participants or only those of a specific age range, race, or religion. In such instances, experimenter control over the characteristics of the sample in the study may be purchased at some cost to external validity (see Chapter 7).

Decide on Sample Size

After identifying a pool of eligible participants, the next decision concerns the *number* of people that will be recruited and included in the experiment. The issue here concerns “statistical power” – the likelihood of obtaining a statistically significant result (if indeed the result ought to be significant) on statistical tests based on **null hypothesis significance testing** (NHST). It is important to ensure that enough participants have been included to detect systematic mean differences between experimental conditions, above and beyond statistical chance. If the number of participants is too low, statistical tests will have lower **statistical power** and may fail to detect a significant difference even though one should be present (a Type II error – see Chapter 2).

Given that sample size matters for statistical testing, how do we know how many participants will be sufficient? The most widely accepted method is to perform a **power analysis** prior to conducting the study. Cohen (1992) provided a useful table for estimating necessary sample sizes for the most common statistical tests. The table is premised on choosing the effect size (with numeric values representing “small,” “medium,” or “large”) that the research would like to be able to detect.¹ A popular free computer software, G*Power (www.gpower.hhu.de), provides a more precise way to determine the sample size requirements. The program requires entering information about the magnitude of the expected effect to be detected, critical rejection region (e.g., $p < .05$, two-tailed), the desired statistical power (80% power is commonly applied), and type of statistical test (e.g., t-test) that will be performed. In a two-group experiment, to detect a “large effect” would require a minimum of 52 participants (26 in each group) to have sufficient statistical power to attain the $p < .05$ level of statistical significance (two-tailed). However, to detect a “small effect,” the sample size would have to increase considerably, to 788 participants (394 in each group) to obtain statistical significance at $p < .05$. Thus, constructing experimental manipulations that are salient and impactful enough to produce larger rather than a smaller difference in group means is strategic in requiring fewer participants for a study. In Chapter 2, we discussed methods of increasing power in addition to simply increasing the number of participants. Effect sizes will be discussed in depth in Chapter 20.

Prepare Materials

Once the participant pool has been identified and the selected sample of eligible participants is defined, the experimenter can organize the various features that will constitute the experiment itself. Instructions must be prepared, independent variables and dependent measures planned and constructed, and the informed consent and debriefings written.

Immediately upon arriving at the experimental setting, participants should be asked to provide their **informed consent**, a process in which the researcher notifies potential research participants of the potential benefits and risks they might experience when participating, so that they may make a reasoned decision about engaging in the study or not. After participants have served in the research, **debriefing** is the experimenter's honest explanation of what the study is about and is especially indispensable if deception was involved (see Chapter 21). Many experimental manipulations can be developed in such a way that it is unnecessary to deceive participants. However, if deception is a necessary feature of the study, the experimenter must develop a strategy to offset its effects in a careful post-experimental debriefing. A good general rule is that participants should depart the study feeling as good about themselves as they did when they entered it. If the experimental treatments put participants at risk, or if they were misled in any way, this potential damage must be offset at the study's completion.

Submit Plan to IRB

Before an experimental study can be conducted, the central features of the experimental design – the instructions, independent and dependent variables, and consent and debriefing messages (the study “protocol”) – must be submitted for approval to a committee specifically constituted to protect the welfare of research participants. In most universities, this ethics committee is called the Institutional Review Board (IRB). It is imperative that no research involving human participation ever be conducted without prior approval of the IRB. Failure to meet this requirement can have serious consequences for an investigator and sanctions on the institution.

Set Up Laboratory Environment

If the IRB has reviewed the protocol and allowed the study to go forward, the researcher then must set up the experimental context. In considering all the elements of an experimental setting, we need to distinguish between those features of the context that are to be *held constant* and those that are to be *systematically manipulated*. Participants must be given some information or instructions regarding the experiment – what it is about and what they are supposed to do. Apart from the specific features that are to be manipulated as the independent variable(s) and the outcome measures constituting the dependent variable(s), it is critical to good experimental design that potential extraneous features of the study be identified and systematically held constant by the experimenter. This is to prevent potential extraneous variables from interfering with or confounding the effect of the independent variable(s) on the dependent variable(s).

Most features other than the independent variable of interest should be controlled for by being held constant to ensure that they are kept the same for all participants in all conditions in the experiment. This makes it possible to draw conclusions that the independent variable is solely responsible for producing differences across groups, as observed on the dependent variable. Ideally, the same laboratory or research space should be used for all conditions, the same set of instructions should be recorded and played for all participants, the same experimenter should administer every session, and so forth. Some features cannot be held constant in this way. It is

difficult, for instance, to administer all sessions of an experiment at the same time of day, and the same experimenter may not be available for every session. In these cases, the researcher should control or hold constant the situation so that contextual variations are evenly distributed across the conditions. Some experimental sessions may be run at 10:00 a.m. and others at 5:00 p.m., but every experimental condition must be as likely to occur at the early sessions as the later ones.

Extraneous features of the laboratory context are usually considered nuisance background factors that must be removed or controlled to avoid confounding with the experimental manipulation (as discussed in Chapter 2). Sometimes, however, a background variable is an integral part of the experimental setup. Consider, for example, an experiment by Latané and Darley (1968) conducted as part of their investigation of “diffusion of responsibility,” the idea that bystanders are less likely to respond to a potential emergency situation if other people are present than if they are alone. To test this hypothesis in a laboratory context, the experimenters arranged a situation in which dark smoke from a ventilator began to pour into the room while participants were completing a questionnaire, either by themselves or with other participants (confederates) present. The number of seconds it took a participant to leave the room to seek help after the smoke appeared constituted the dependent variable. As predicted, the more people in the room, the longer it took a participant to seek help.

In this example, the environmental feature of smoke pumped into the room constituted a part of the setup or staging of the experiment. It was held constant for all levels of the independent variable – in other words, the same amount of smoke was piped regardless of whether each participant was assigned to the room to be alone or with others. The independent (manipulated) variable was the number of other people present but the amount of smoke was held constant across the conditions.

Types of Experimental Manipulations

We classify experimental manipulations into five broad types: **environmental manipulations**, **stimulus manipulations**, **social manipulations**, **instructional manipulations**, and **priming manipulations**. These manipulation varieties may be combined in the same study, but for illustrative purposes we will discuss them separately as “pure” types.

Environmental Manipulations

Environmental experiments involve a manipulation of the properties of the physical environment or setting of the experiment. In these types of manipulations, features of the environment, such as lighting, background noise, color of the room, objects in the room, or room temperature, are systematically varied by the experimenter to determine whether the manipulated context or environment leads to differences in how participants respond on the dependent variable.

An example of an environmental manipulation involves modifying room temperatures to observe its impact on aggression and hostility (Anderson, 1989, 2001). Experiments that modify room temperature to elicit anger are based on the theoretical perspective that excessive heat makes people feel uncomfortable and distressed. In terms of underlying mechanism, the external discomfort of heat experienced by participants is misattributed as internal discomfort that in turn leads to coping with the heat through aggressive and hostile reactions. In typical experiments involving the environmental manipulation of temperature, participants are randomly assigned to a hot room (e.g., 95 degrees Fahrenheit) or a comfortable room (e.g., 76 degrees Fahrenheit). While seated in each room, participants are asked to complete dependent variables such as self-reported state of aggression, hostility, negative mood, and stress. In other studies

involving different operationalizations of the dependent variable, participants might engage in a task to test aggressive actions and their behaviors may be observed by coders through a one-way window. Some experiments such as these have supported the hypothesis that exposure to a hot room compared to a normal temperature room elicits stronger anger tendencies and reactions.

In another interesting use of “background” environmental features, Berkowitz and LePage (1967) exposed participants to either one (minimum) or seven (maximum) electric shocks in the first part of the experiment. In the second part, the participants were given the opportunity to retaliate by administering shocks to the person (confederate) who had previously shocked them. Objects sitting on a table near the shock contraption were varied as an experimental manipulation. In one condition the table held a 12-gauge shotgun and a .38 revolver; in other conditions the objects were two badminton racquets with shuttlecocks, or no objects at all. In all conditions, participants were told to ignore the objects as irrelevant (leftover from other experiments). Nonetheless, results indicated that participants who had previously received maximal shocks in the weapons condition delivered significantly more electric shocks to their partner than those in the badminton or no weapons conditions. The mere presence of the aggression-arousing environmental cue of weapons elicited greater retaliatory aggressive behavior from the participants.

Stimulus Manipulations

Stimulus manipulations involve the explicit presentation of materials that involve text (vignette), visual, or auditory information (or their combination), to which participants are to respond in some way. A large proportion of social psychological experiments involve the use of stimulus manipulations to study either how people make judgments and decisions, or how they are influenced by persuasive communications.

Judgment experiments. Typically, in studies of judgment and decision-making, participants are given some visual or verbal materials and asked to make judgments, attributions, or decisions about the content of those materials. In such stimulus manipulations, the independent variable consists of variations in aspects of the stimulus materials that are presented. For instance, in much of the experimental research on impression formation, or person perception, participants are given a verbal description of some individual person (sometimes with a photograph accompanying the description) and are asked to answer questionnaire items about how much they like this person or what traits that person might possess. In these studies, specific pieces of information about the person depicted in the stimulus materials are manipulated (e.g., age, ethnicity, physical or psychological traits, specific behaviors) to determine how those manipulated features influence participants’ judgments.

Attribution experiments, in which the independent variable involves features of the stimulus or information given to the participant to respond to, are used across a wide array of research topics. For example, in criminal justice research, experimentally manipulating the description of factors embedded into a police report, such as the type of crime committed (rape or indecent exposure) or whether the offender was intoxicated or not, is used to determine their effect on participants’ ratings of guilt or innocence.

In one example of an attribution study applying a stimulus manipulation, participants evaluated a described person in terms of a perceived mental health stigma (Correll et al., 2021). All participants read a vignette involving a paragraph of text that described the functioning of a person (“Sammy”) with a post-traumatic stress disorder (PTSD) diagnosis and symptoms, such as waking up from nightmares about the trauma, flashbacks of the traumatic experience, and spacing out. Participants were randomly assigned to a photograph manipulation of the same person wearing either a civilian outfit (casual dress shirt) or military outfit (army combat uniform).

All other materials were held constant in both conditions, including the name of the person, the text describing the PTSD symptoms, and the photograph of the person showing a neutral facial expression. The photograph was embedded next to the text in the vignette. For the dependent variables, participants rated the person on various attributes related to mental health stigma. The results revealed that the military service member, compared to the same person as the civilian, was rated as more dangerous, fear-inducing, and needing to be segregated from society.

Communications and attitude change. In another body of research, stimulus materials are manipulated to test their effectiveness in creating or altering participants' beliefs, attitudes, or behaviors. In this genre, we are not interested in how participants judge the written or oral communications but how the content or style of those messages influences their perceptions of the topic of communication (e.g., policies, products, ideas, groups, etc.) or their future behaviors. Manipulations of stimulus materials include varying the information about the source of the message, the content of the message itself (e.g., the strength of arguments), or some combination of these. A huge body of research on communication and persuasion documents the impact of message content and source (Crano & Prislin, 2006; Tormala & Rucker, 2022; Verplanken & Orbell, 2022).

Another feature of a communication that may be manipulated is the apparent intended audience to which the communication is directed. In research on prevention of inhalant abuse in young adolescents, for example, Crano et al. (2007) played a persuasive anti-drug video to sixth- and seventh-grade school children. For the *indirect message* group, the video began with a speaker who said, "Parents, do you have a young teen at home?" In the *direct message* group, the video began with "Are you in the sixth, seventh, or eighth grade?" Both videos went on to discuss the dangers of inhalant use. Students exposed to the persuasive indirect message designed for their parents were considerably more persuaded by the message than those who were the obvious direct targets of the message. It seems that participants exposed to the indirect message compared to the direct message did not feel as strong the need to defend themselves against the video message.

Social Manipulations

Social manipulations involve investigating the presence or absence of other people and their behaviors on the responses of participants. In many social manipulations, for the purpose of being able to experimentally manipulate the social influence of others, these other people are confederates who are actors who adhere to a planned script while pretending to be another participant in the study. As social manipulations tend to involve confederates posing as other participants, these designs almost always involve some degree of deception.

The use of a social manipulation is exemplified in Asch's (1948, 1951) classic studies of conformity, in which a naïve participant was placed in a room with confederates. Participants were told a "cover story" that the study was designed to investigate perceptual processes. On each trial, a large card containing a standard line was presented along with three comparison lines. The participant's task was to judge which of the three comparison lines most closely matched the standard line in length. The three comparison lines differed enough in length so that the correct choice was obvious and did not demand fine visual discrimination skills (see the illustration in Figure 6.2). The session was set up so that on each trial the naïve participant responded after the confederates; the judgments given by the confederates were scripted so that on some of the judgment trials, they unanimously chose a clearly incorrect comparison line as the response (e.g., naming comparison line #3 in Figure 6.2 as the match instead of the clearly correct line #2).

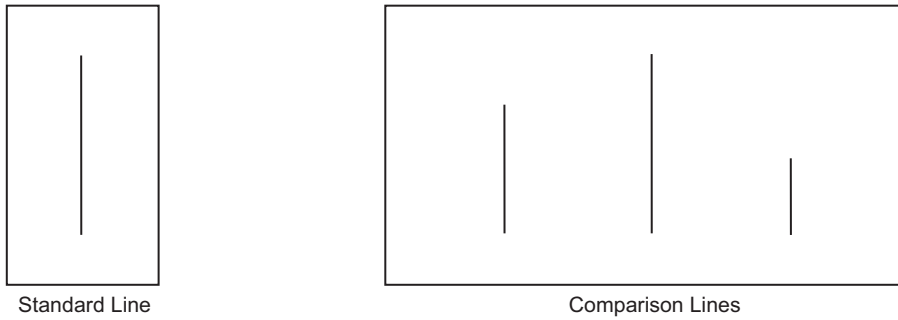


FIGURE 6.2 Example of standard and comparison lines used by Asch (1948).

Asch's experiments were conducted to determine how behavior on the part of other people in the setting would influence the overt choices made by the naïve participant. Results indicated that participants could best resist the group's influence of making an incorrect judgment in this seemingly easy perceptual task when only one or two confederates were present; the addition of a third confederate dramatically increased the degree of conformity to the incorrect answer, but further increases in group size had no substantial incremental effect. In some of Asch's experimental conditions, as many as 15 confederates were employed for each participant!

Other social manipulations may not employ in-person confederates but rely on virtual and electronic confederates for cost saving and better standardization across conditions. With technology, these simulated social interactions can be made to appear quite realistic. For instance, in experiments conducted by Lord and Saenz (1985), the researchers placed participants into a social situation in which they believed they were "solos" – the only woman (or man) in a group of men (or women). Actually, all the other people were actors whose responses were prerecorded, with the recordings played back during the appropriate times by the researcher. The real participants were placed in a small room facing a video camera and were led to believe that they were interacting with other participants via "live" video broadcasting. To make the experience as authentic as possible, the participants were recorded and saw themselves on a video monitor whenever it was their turn to speak. This feedback made the manipulated social interaction quite compelling, and no participants reported being suspicious about the live presence of other participants (who were confederates) when they were later debriefed about the actual purposes of the study. Similar elements of realism have been used in simulated interactions on computer "chat rooms," where the actual participant was led into believing that the other people were interacting live in the electronic room, but these were actually simulated messages of people pre-scripted by the experimenter (Gardner et al., 2000).

Instructional Manipulations

Another type of experimental manipulation is one that involves modifying instructions, directions, or procedures that the participant is to follow. In the typical instructional manipulation, the instructions or procedures provided to the different groups are identical except for the substitution of a few words or sentences. Thus, it is extremely important that participants are paying attention to reading or listening to the instructions, so they "receive" the appropriate experimental manipulation.

Instructional manipulations can be embedded in the introductory materials that are presented in an experiment. For example, an experiment by Brewer et al. (1993) used the wording

of the confidentiality instructions to manipulate whether participants would feel highly individuated (personalized) or depersonalized within the experiment. In the depersonalized condition participants were told that they had been assigned an arbitrary ID code and the wording emphasized that the research was interested in “the attitudes and perception of students in general” and that “you represent an example of the average student.” In the more personalized (control) condition, standard confidentiality instructions were used and participants were able to generate their own individual ID code. These differences in the introductory wording proved to be sufficient to influence participants’ later responses to being assigned to majority or minority groups.

Even small differences in the wording of instructions and directions can prove to have powerful effects. Consider, for example, the research by Zanna and Cooper (1974) who were studying the effects of misattribution of arousal on people’s judgments. Zanna and Cooper gave all participants a placebo pill (which consisted solely of milk powder), but the instructions were manipulated to indicate different expectations about the pill’s supposed effects. Participants who were randomly assigned to the “arousal” treatment condition were told:

This M.C. 5771 capsule contains chemical elements that are more soluble than other parts of the compound. In this form of the drug these elements may produce a reaction of tenseness prior to the total absorption of the drug, 5 minutes after the ingestion. This side effect will disappear within 30 minutes.

(Zanna & Cooper, 1974, p. 705)

Participants who were assigned to the “relaxation” condition were given the same pill, except that the instructions involving the word “tenseness” was replaced by “relaxation.” This simple manipulated variation in wording of the instructions established different participant expectations and interpretations of their own bodily reactions after ingestion, which, in turn, altered their response on the dependent variable.

Another form of instructional manipulation involves variation in the task that participants are given to do, or the way that the task is described. In early experiments on social perception, for example, participants were given a series of behavior descriptions to read. Prior to seeing the stimuli, participants were given one of two different task instructions that were intended to influence their processing goals. In the *memory condition*, they were told to read the descriptions with the purpose of remembering them as well as possible in a later memory test. In the *impression formation condition*, participants were told to read the sentences with the goal of forming an impression of a person who had performed those behaviors. Although the information that participants then read was the same in both conditions, the processing instructions resulted in differences in what participants remembered in a later recall task. Participants in the impression formation condition recalled more behaviors and also clustered the remembered statements differently in their recall protocols compared to memory condition participants (Hamilton et al., 1980). Thus, the processing instructions affected not only what was remembered but how that information was organized in memory.

Priming Manipulations

Priming manipulations involve inducing a particular mindset or mentality designed to carry over and influence how later information is interpreted or experienced (Bargh & Chartrand, 2000). The main objective of a priming experiment is to activate a mental representation (also known as a “schema”) or trigger psychological preparedness in participants. These induced mental representations could involve thoughts, goals, and expectations. Priming experiments

are best described in terms of two phases. In the “priming phase,” participants are exposed to or engage in a priming task to activate a mental representation. In the “testing phase,” the carry-over impact of the mental representation on dependent variables is assessed.

In designing a priming manipulation, it is important to ensure that participants are unable to fully recognize that the purpose of the priming phase is to influence their subsequent responses on the testing phase. Otherwise, participants might retaliate by compensating for or nullifying the effect. To prevent this from occurring, the design and procedures should be inconspicuous so that participants do not fully recognize that the mental state activated by the priming phase is intended to impact scores during the testing phase. One way of doing this is to present a “cover story” to induce participants to believe that the priming and testing phases are not connected or part of the same study. For example, after the priming phase, participants might be asked to go to another room for a “different study” in which the testing phase is presented. A second possibility is to administer a distractor activity (e.g., connect the dots puzzle) between the priming phase and testing phase.

An example of a priming manipulation is a task used by Brewer and Gardner (1996) to activate thoughts about the self as an individual, or the self as a part of a group. During the priming phase, in the guise of a proof-reading exercise, participants were given the task of circling all the pronouns in an extended passage of a paragraph they were to read. Two different versions of the passage had been created by the experimenter. Half the participants read a version in which almost all the pronouns were the words “I,” “me,” or “my.” The other participants read a version of the identical passage except that the pronouns were all “we,” “us,” or “our.” In a testing phase of the experiment, participants completed a “Who am I?” task in which they were to generate up to 20 aspects that described themselves. Responses to the self-descriptions were analyzed to determine whether the previous activity had influenced whether participants primarily thought about themselves in terms of personal traits or social relationships.

In another example of priming, Carter et al. (2011) assessed participants’ attitudes and voting intentions toward presidential candidates (Barack Obama and John McCain) immediately prior to the 2008 election. All participants received the same questionnaire containing the same set of questions. Participants were randomly assigned to either the condition in which a small American flag was printed or not printed on the top left corner of the questionnaire. The flag was inconspicuously placed at the corner and participants were not asked to pay attention to it or asked questions about it. The findings showed that the presence compared to absence of the American flag caused an increase toward Republican and conservative leaning voting intentions, voting behaviors, and political beliefs, with some of these results lasting eight months after exposure.² This subtle manipulated presence of the flag (priming phase) was demonstrated to sway responses to political attitudes and intentions (testing phase).

Items in a questionnaire could also serve as the priming task in the priming approach known as *questionnaire priming*. Lac and Donaldson (2021) presented questionnaire priming stimuli regarding the norms of alcohol attitudes (peer injunctive norms) or norms of alcohol behaviors (peer descriptive norms) to influence personal alcohol beliefs and behaviors in college students, who were randomly assigned to one of three conditions: peer descriptive norms priming, injunctive norms priming, or control group that did not receive any priming questions. During the priming phase, participants in the peer injunctive norms condition were primed with a series of questions such as “Do your friends approve of drinking alcohol?” and those in the peer descriptive norms condition received questions such as “How often are your friends drinking alcohol?” During the testing phase, the dependent variables included personal attitudes, motivations, and behaviors toward alcohol. The priming phase and testing phase were embedded into a single self-report questionnaire so that participants would not realize that this was an

experimental study in which the former was supposed to influence the latter. The findings tended to support that the priming of social norms about alcohol in college students in both experimental conditions compared to the control condition, which produced a peer pressure effect in generating higher endorsement of the personal alcohol measures.

A particularly interesting type of priming manipulation involves tasks that are designed to create specific bodily experiences that are hypothesized to influence (or prime) related mental content. The ideas underlying such priming manipulations come from theories of *psychological embodiment* – the idea that certain bodily states or physical actions are closely associated with psychological states such as evaluative orientation or emotions (Niedenthal et al., 2005). Thus, engaging the bodily state can prime or activate the corresponding psychological state. In an early example of psychological embodiment, Wells and Petty (1980) demonstrated that randomly assigning participants to either nod or shake their heads (priming phase) while listening to a persuasive communication subsequently influenced the extent to which the participants accepted the arguments presented in the persuasive communication (testing phase). Those who were required to nod their heads up and down compared to those who shook their heads left and right while listening to a communication were more persuaded by the message.

Manipulation and Attention Checks

A **manipulation check** is a question (often asked at the end of an experiment) that assesses whether participants noticed or picked up on the manipulation they received in the experiment. Participants' level of engagement in the study is a critical factor in determining whether the independent variable was effective in impacting scores on the dependent variable. The manipulation check serves as an opportunity to determine if the manipulation in their assigned condition was salient enough for participants to detect. Suppose, for example, that participants were randomly assigned to one of three conditions in which they read a police report depicting a perpetrator of a crime who was intoxicated with alcohol, marijuana, or cocaine. The manipulation check question might be "What substance did the perpetrator use?" with the multiple-choice options of alcohol, marijuana, or cocaine. Participants who answered the question correctly, based on their assigned condition, passed the manipulation check.

Manipulation checks may also assess whether the participants perceived the manipulation as intended. Suppose, for example, we were interested in determining the effects of fear-arousing safe sex advertisements on the at-risk sexual behavior of adolescents. After exposing some participants to ads with fear-arousing content and some to control ads with no fear content, we monitor their reported behavior over the next six months. If no differences in reported behavior were found between those exposed to the ad campaign and those who were not, we might be tempted to conclude that fear arousal does not affect the sexual behavior of adolescents. We might be correct in this interpretation, but it might rather be the case that the ads in the fear condition were not all that fear-arousing. If we had administered manipulation checks after the ads were read – such as "How frightening are these ads?" – we could at least determine whether the ads in the fear-arousal condition scored higher on this manipulation check question compared to the control group.

When most participants pass the manipulation check, or rate the manipulation as intended, we often say that the manipulation was "successful." However, it is important to keep in mind that this simply means that the manipulation was received as intended. Manipulation checks do not tell us whether the manipulation *affected* the participants in the way that we hypothesized. The effects of the manipulation are determined by the outcomes on our dependent measures (or process measures that might be included in the experiment). Manipulation check information

can be useful to the experimenter, even when the experimental treatment has apparently failed. Nonetheless, we need to caution that such checks – while potentially informative – are not necessarily perfect indicators of whether our experimental manipulations have been received as intended. Participants may not always be able to state explicitly what they have experienced or understood, particularly when a self-report measure is used for the manipulation check (Nisbett & Wilson, 1977). Thus, manipulation checks can be used in conjunction with the results on the dependent variables in an experiment to help understand the findings, but they should be used only as an aid in the process of interpretation.

An **attention check** is a question inserted within the experimental materials as a test to determine whether participants are attentively reading the instructions, stimulus materials, and/or questionnaire items in a study. Some studies use attention check questions in only a certain part of the questionnaire (with the idea that if participants didn't pay attention to that part, then they likely didn't pay attention to the rest of the questionnaire), whereas other studies intersperse different attention check questions throughout the study materials. An attention check question might ask participants, "Please enter the number 9 for this question," with participants who enter a different value or provide a blank response not passing the attention check. In another example, participants might be asked, "Have you ever died from hunger?" with individuals answering "yes" not passing the attention check.

The distinction between a manipulation and attention check is that a manipulation check verifies whether participants paid attention to the independent variable and therefore perceived or experienced their assigned experimental condition, whereas an attention check verifies whether participants paid attention to all other aspects including the instructions and dependent variables in the study. As such, an experimental study might include manipulation check and/or attention check questions, whereas a nonexperimental study might include attention check but not manipulation check questions. Manipulation and attention checks should be used sparingly to screen participants, as the overuse of verification check questions could make participants suspicious of the researcher's intentions and violate their trust.

Responses to both manipulation and attention checks can be useful to help researchers know whether their experimental materials are "working." If all or most of participants pass these checks, confidence in the data resulting from the study can be enhanced. However, the question remains what to do about participants who do not answer the manipulation check "correctly" or who fail the attention checks. Some studies delete all participants who do not pass all the manipulation and attention check questions before performing statistical analyses, to yield a cleaner dataset. Other studies report the percent who don't pass the checks but choose to keep the data from those participants to maintain the entire sample.

Assignment of Participants to Conditions: Randomization Procedures

We have discussed the importance of random assignment in developing true experimental designs that allow for causal inference. True random assignment requires that any person in our sample be equally likely to participate in any of the experimental conditions. This is the central requirement of random assignment, and without it we cannot claim that our assignment was random. Various ways are available to implement random assignment once experimental conditions have been designed and constructed, so we cannot go into extensive detail about all those methods here. However, we can present a few basic principles about the mechanics of random assignment.

In general, two different ways are available to accomplish random assignment of individual participants to experimental conditions. One involves using some process of randomization (e.g., a coin toss, roll of a die, or use of a table of random numbers), one-by-one as each

participant arrives at the experimental session. For instance, if we have a 2×2 factorial design, resulting in four different experimental conditions, we could label the various conditions 1, 2, 3, and 4 using four different slips of paper and then place them in a bag. As each participant shows up, we would randomly draw a number from the bag and whichever number comes up will be the condition to which the participant is assigned. The drawn number would then be returned to the bag before drawing another number corresponding to the condition to which the next participant should be assigned. This method of randomly assigning participants on a one-by-one basis is the form of random assignment procedure called *simple random assignment*. However, it has a disadvantage: we must run many participants to be sure that the resulting randomization will spread the same number of participants evenly across all our conditions. With simple randomization, we can by chance draw a long string of 1s all in a row, so that participants get “bunched up” into one condition until a large enough sample size will even things out due to chance. This is not particularly problematic if we have many participants in our study. In that case, this randomization process will work quite well in distributing approximately equal numbers of participants across the conditions. However, if the study does not have many participants available, this random assignment procedure may not work so well and we wind up with unequal *ns* across conditions.

The alternative form of random assignment is *block randomization* in which random assignment occurs for each block of participants until all participants are assigned to conditions. All possible conditions are numbered on slips of paper and then randomly ordered, and then this random ordering is repeated in blocks until the total number of intended participants is reached. For example, using our previous 2×2 design, the four conditions representing numbers 1 to 4 would be placed in a bag. Then, the first randomly drawn number would indicate the condition administered to the first participant (the bag will now have three slips remaining), the next number drawn would be the condition administered to the second participant (the bag will now have two slips remaining), and so forth. Once all four conditions had been included (no more slips in the bag), the process would be repeated for the next block of participants, and so on. If we were planning to run 25 participants per condition, for a total of 100 participants, this process of randomly ordering the four conditions would be done 25 times (representing 25 blocks). This process of blocked randomization assures that participants are both randomly assigned *and* evenly distributed across the four conditions.

Technically, random assignment should be implemented in such a way that each individual participant in the experiment has an equal chance (at the outset) of being assigned to any of the conditions, independently of any other individual's assignment. Both simple random assignment and blocked random assignment procedures meet this criterion, as long as participants are allocated separately to their randomly determined condition. However, it sometimes is the case that a particular experimental condition is delivered to all participants who show up at a particular session at the same time. This is called *group randomization*. For instance, we may have one independent variable in which an instructional manipulation is to be delivered verbally by the researcher. At the same time, we have a second independent variable involving a stimulus manipulation that is contained in written materials distributed individually to each participant. Now say that we are conducting experimental sessions in which five participants are studied at the same time in the same room. We can still randomly assign each of these five participants to different levels of the written stimulus materials because each participant does not see what the others have received. However, everyone hears the same verbal instructions read aloud by the experimenter, so only one version can be delivered. In this case, the levels of the instructional treatment are randomly assigned to whole sessions, with all five participants in the same session getting the same version.

This process of group random assignment is implemented *across* sessions, but not *within* sessions. Technically, group randomization is a violation of the principles of random assignment at the individual level because there may be nonindependence among participants in the same session, which must be statistically controlled for (see Chapter 19 for discussion of how this affects statistical analyses). Nonindependence in each session occurs because those who choose to appear at the same session may possess similar background characteristics or preferences (e.g., people who possess certain traits may prefer the morning sessions), or the room conditions at the time of the session may affect all participants in the same way.

If we run many sessions across the course of an experimental study, this violation of random assignment is not particularly problematic, but it can lead to loss of data at times. For example, suppose we want to study the effects of discomfort on people's judgments of various governmental policies (e.g., welfare, job training, universal health insurance). We have a laboratory that holds 20 participants at a time and seek to administer the same treatment to everyone in the room. We want to make some of them uncomfortable to determine whether discomfort will make them less likely to support government-funded welfare, job training, and health insurance. To do this, we play loud white noise during the "discomfort" condition. The controls are studied under normal circumstances. If we wanted to randomly assign 120 participants (60 into each of the two conditions), we would need to conduct six sessions – three treatment sessions and three control sessions. We want the sessions to be identical in all respects *except* the presence or absence of the treatment. However, suppose that during one of the sessions, one of the participants became very ill and suffered a seizure. This disruption probably would render the data of this session noncomparable with that of the other sessions, and we would probably have to drop the data from that session, losing all 20 participants. For this reason, it is a good general rule that if experimental treatments are assigned to sessions, one should conduct many sessions, with only a few participants (e.g., no more than five) in any one session.

Realism and Engagement in an Experiment

The frequent use of subtle manipulations as independent variables in experiments highlights the importance of participant involvement and attention as a critical factor in good research. In this connection, Aronson et al. (1968) made an important distinction between experimental and mundane realism, concepts that are closely tied to the way experiments are set up and the types of manipulations used. **Experimental realism** is the degree that an experiment has real impact on participants as they are engaged in the study. To achieve experimental realism, the experimental arrangements are designed to ensure participants attend carefully to the situation and the task at hand. In other words, experimental realism is achieved if participants do not over-intellectualize and are not self-conscious about their reactions but are responding to the experimental situation in a way that approximates their natural and spontaneous behavior outside the laboratory.

Mundane realism is the degree to which various features of the experiment – instructions, treatments, and measurement operations – mirror real-world, nonlaboratory events that participants might encounter in their day-to-day lives. For instance, some experiments involve asking college students to write essays, or take tests, events that are quite usual in the daily life experiences of the average student. Note, however, that this same experimental task would have considerably less mundane realism for participants who were middle-aged truck drivers for whom such tasks might be considerably more exotic.

Mundane and experimental realism are not mutually exclusive. Whenever possible, a good research design will establish both. However, of the two, experimental realism is the more

important for creating a meaningful and engaging situation for research participants. The mere fact that an event occurs in real life does not endow it with importance, and an experimental situation that mirrors a dull real-world experience will probably prove to be dull and uninvolving. Research participants in this type of setting will tend to become apathetic and may fail to respond to the manipulation simply because they did not attend to it. Some of the treatments used by researchers are so subtle that participants must be closely attuned to the research situation if the manipulation is to have any impact on their responses. If respondents are bored or apathetic, this needed degree of attention cannot be assumed. It is clearly worth the time and effort that it takes for researchers to keep their participants interested and involved.

Role-Playing Simulations and Analogue Experiments

The techniques for conducting laboratory experiments discussed in the preceding sections of this chapter emphasize achieving experimental realism within the laboratory setting. A major goal is to design experimental procedures to create an environment that is engaging and impactful for research participants, even if the situation is peculiar to the laboratory setting and bears no direct resemblance to events the participant is likely to encounter in life outside the laboratory. Now we will shift attention to laboratory studies with the explicit purpose to emulate events that occur in the “real world.” In such studies, the degree of correspondence between the situation created in the laboratory and the real-life situation it is intended to represent becomes a major focus of how the research is designed and conducted.

Two types of laboratory experiments share this concern for real-world correspondence. The first is the **role-playing simulation**, in which participants are instructed to imagine or play a role in a specified real-world situation or scenario and to respond as they believe they would in that context.³ Role-playing simulations could be further subdivided into two types. In a *passive role-playing simulation* participants imagine and visualize the entire scenario created by the experimenter and their role in it, but do not physically act it out. Instead, they report how they believe they would behave or respond in that situation. In an *active role-playing simulation*, based on the role and scenario assigned by the researcher, participants physically act out the entire scenario.

The second type of laboratory research concerned with real-world correspondence is referred to as the **analogue experiment**, in which participants are responding directly to a specially constructed situation that has been designed to reproduce or mimic selected features of a real-world situation. An analogue experiment differs from a role-playing simulation in that participants are not asked to play an explicitly predefined role and therefore their behavioral responses are natural. Another distinction is that analogue experiments compared to role-playing simulations sometimes involve some form of deception. In a role-playing simulation, as participants are given an explicit role and scenario, they are aware of the purpose of the study. In an analogue experiment, the deception of participants is often part of the design, so the natural reactions of participants can be observed, because otherwise participants might react in an unnatural manner.

Role-Playing Simulations

When used for research purposes, simulations are intended to preserve many of the advantages of controlled laboratory experiments while approaching features that are more generalizable to the real world. A well-designed simulation has the potential to isolate the social phenomenon of interest without destroying its natural contextual meaning because participants can “actively imagine that all the normal constitutive relations of a social situation are satisfied” (Greenwood,

1983, p. 243). Because of these added “imaginative” elements, the treatment conditions of a simulation study are inevitably more complex and multidimensional than those of the basic laboratory experiment. Hence, the potential increase in generalizability to the real world is attained with some sacrifice of precision in the specification of the independent and dependent variables.

Passive role-playing simulations. Various types of simulation research differ in the extent the role-playing is passive or active. At one end of the spectrum are studies employing what Greenwood (1983) calls “passive interpretive role-playing,” which might also be described as “mental simulations.” In such studies participants are provided a written or verbal description of a situation or scenario, and their role in it, and are asked to estimate or predict by imagining how they (or others) would behave in that situation.

Such role-playing studies have been used for theory-testing purposes on occasion. Rosenberg and Abelson (1960), for example, placed participants in a role-playing situation to test some hypotheses derived from theories of attitudes and attitude change. In their study, participants were asked to imagine themselves in the role of a department store manager. As part of the context of their role-playing, participants were “given” a set of attitudes toward a particular employee (Mr. Fenwick) and his plans to mount a modern art display in the store’s rug department. The interpersonal relationships assigned to different participants were manipulated to produce different states of psychological inconsistency or imbalance. Participants were then given three different communications to read, each of which was designed to change their attitude toward some aspect of the situation. A dependent measure of final attitudes was used to determine which communication was accepted by the role-playing participants and the extent this corresponded with researcher predictions derived from formal balance models. Although the results confirmed balance theory predictions to some extent, they also indicated that acceptance of communications was impacted by motives other than restoration of cognitive balance (such as avoiding negative interpersonal emotions). Thus, even though it did not involve a “real” interpersonal situation, this passive role-playing study did prove useful in testing the theory.

Some forms of passive role-playing have been suggested as possible alternatives to the use of deception in experimental psychology (e.g., Carlson, 1971; Kelman, 1967; Mixon, 1972; Schultz, 1969). The argument here is that if role-playing participants can be given a complete subjective understanding of the experimental conditions that would be used in a real study, their envisioning of how they would behave in that situation can substitute for actual participation in such situations. During the 1960s and 1970s, many role-playing studies were conducted to determine whether the results obtained would match or reproduce the findings from previously conducted deception experiments. Horowitz and Rothschild (1970), for example, compared reports from two forms of simulation against the data from an earlier Asch-type conformity study conducted by Gerard et al. (1968). Bem (1965, 1967) conducted a series of “interpersonal simulations” of classic dissonance experiments (Brehm & Cohen, 1959; 1962; Festinger & Carlsmith, 1959), and the outcomes of Milgram’s (1965) studies of obedience have been subjected to role-playing simulation (Freedman, 1969). In terms of their ability to reproduce findings in which participants are to be themselves rather than playing an assigned role, the results have been mixed (see Miller, 1972). Even when the findings are parallel, it remains ambiguous whether the results of a passive role-playing simulation can be interpreted in the same way as those obtained under the real experimental conditions (see Cronkite, 1980).

Active role-playing simulations. Although passive role-playing has some potential value for theory testing, the more widely used forms of simulation involve active role-playing efforts in which participants are allowed to act out their natural responses based on their interpretation of their assigned role in the simulated social situation. In one classic study in the attitude change literature, Janis and Mann (1965) used active role-playing as a technique to change behavior.

Women who were smokers were asked to pretend that they had just received a diagnosis of lung cancer and were asked to act out the emotions they would feel and how they would tell their significant others about the cancer. Women who had gone through this role-playing exercise were smoking significantly less six months later than women in a control condition. In this case, role-playing was an experimental manipulation. In other role-playing experiments, however, it is the participants' behavior during role-playing that is of interest.

The primary version of this research method is the so-called role-playing game. In this form, participants are given roles to play within a specified social system. The parameters of the system are under the control of the experimenter, and within this context the participants make choices and decisions befitting their perception of the roles they have been given. The behavior choices of each participant and their consequences for the behaviors of other participants in the system constitute the major dependent variables in this research design. Participation in such games is usually extended over a considerable period, and experience indicates that motivation and involvement among role players run quite high. Elaborate role-playing simulation exercises have been developed, ranging from simulations of small decision-making groups, such as jury deliberations (Breau & Brook, 2007; Cox et al., 2013) or management and production teams (Adobor & Daneshfar, 2006; Xu & Bernard, 2013), business organizations and market economies (Galtier et al., 2012; Klein & Fleck, 1990), to whole societies (SIMSOC: Gamson, 1969; Gamson & Stambaugh, 1978) and intercultural relations (Fowler & Pusch, 2010; Hofstede & Murff, 2012). There is also considerable research on the decisions and behavior of individuals in two-person bargaining and negotiation games, a form of active role-playing simulation where participants are asked to adopt roles as negotiators or bargainers in a business or political context.

Probably the most dramatic simulation of a social subsystem is represented by the prison simulation designed and conducted by Philip Zimbardo and his colleagues at Stanford University (Zimbardo et al., 1973). In this active role-playing simulation, Zimbardo created a mock prison in the laboratory basement of a college building and recruited college student participants who were randomly assigned to play the roles of "guards" and "prisoners" and to act out their behaviors in the simulated setting. Zimbardo was attempting to demonstrate the powerful effects of institutionalization and deindividuation in the prison system by being assigned an institutional role (guard or prisoner) on interpersonal behavior. As with Milgram's (1963) earlier studies of obedience, however, the results of the simulation were more extreme than expected as the participants became fully immersed in their respective roles. Although the simulation had been intended to extend across a two-week period, Zimbardo felt forced to cancel the study at the end of six days because of the escalating cruelty on the part of the guards toward the prisoners, who were showing signs of progressive apathy and depression (Zimbardo, 2007).

Analogue Experiments

In role-playing simulation research, the "real world" is represented in the laboratory through the researchers' instructions to role-playing participants and the players' ability to imagine themselves in the situation portrayed. The logic of an **analogue experiment** is quite different in that participants are not asked to act out a role in an "as if" context. Rather, they are presented with a *real* situation to respond to directly and naturally in an analogue experiment. The difference between an analogue experiment and other basic laboratory experiments is in the design of the stimulus situation and its relation to some specified social event or problem existing outside the laboratory. An analogue is designed to preserve an explicit relation between the laboratory setting and some real-world situation of interest; for every feature of the external situation that is considered theoretically relevant, there is a corresponding feature contained in

the laboratory situation. In this sense, an analogue experiment is like a roadmap of a particular geographical region, where there is a one-to-one correspondence between features on the map and essential features of the actual terrain (e.g., highways, rivers, mountains, etc.), but where less relevant features that exist in the real setting (e.g., trees, houses) are not represented on the map. If the important features represented in the analogue situation have been appropriately selected for the laboratory, participants' responses to that situation should provide an accurate "mapping" of their responses to the corresponding situation in real life.

Analogue experiments have been used for some time. Animal models, for instance, are regarded as analogues to human physiology in much medical research, and experimenter-participant roles are treated as analogues to the doctor-patient relation in both medical and clinical research. The use of analogue experimentation in research, however, has been relatively rare, despite its potential role in closing the gap between research in the laboratory and in the real world. Partly, this may be because analogue experiments are not easy to design, as they require concern for both experimental realism and accurate simulation of mundane events. By calling attention to specific examples of such applications, we hope to increase awareness of the analogue experiment as a potentially valuable research tool.

One prototypic example of the use of analogue experiments in social psychology was prompted by a specific real-world event – the killing of Kitty Genovese outside a New York apartment complex in 1964. In the middle of the night, Ms. Genovese was pursued and murdered by her assailant on the streets over a period of over 30 minutes, even though her cries for help could be heard by as many as 38 people in nearby apartments. This event, and the extensive newspaper coverage it generated, led researchers Latané and Darley (1968; 1970) to speculate about the psychological processes involved in an individual's decision whether to intervene in such emergency situations, and how those processes might be affected by the actual or inferred presence of other people who fail to help. Their studies led to the development of a theory of the bystander effect, which postulated that the greater the number of people who did not react to an emergency, the more likely any individual person also would conform by not reacting to that emergency, due to the "diffusion of responsibility."

Among the studies of the bystander effect, the experiment most closely analogous to the Kitty Genovese news event was that conducted by Darley and Latané (1968). In this experiment, a participant was given the cover story that this was a communication study that required them to sit in a separate small booth to communicate via audio intercom with the other "participants." Only one real participant was tested in each trial. The other communications were scripted tape recordings of other ostensible participants. After a brief warm-up introduction, each person was asked to speak, one at a time, about their college experiences. Midway through his presentation, in which he had mentioned his history of epileptic seizures, one of the apparent participants (a recording prepared by the experimenters) said he felt lightheaded. His speech became blurred and disoriented, and he eventually was heard audibly gasping for help. After this, a thud was heard over the intercom and then silence, suggesting the victim had fallen to the ground.

The independent variable in this experiment was the number of passive participants apparently engaged in the study at the time of the emergency. In one condition, participants believed that they and the "victim" of the seizure were the only two participants in the session; in the other two conditions, each participant believed that in addition to the victim there were either one or four other bystanders present in the experiment. As in the real-world event involving the Kitty Genovese case, the actual participant was physically isolated from the potential victim and from other "bystanders," and, as in the real emergency, the participant had to decide how quickly to respond to the emergency. The dependent variable was the time it took the naïve participant to leave the booth and report the emergency. When participants believed they were

the sole potential helper in the situation, 85% responded within less than a minute to seek the assistance of the nearby experimenter. When participants believed others were present, the likelihood of responding within a short interval dropped dramatically as the number of bystanders increased. Only 31% responded when they believed there were four other bystanders available. Known as the “bystander effect,” this phenomenon of being less inclined to offer help when passive others are present is attributed to diffusion of personal responsibility.

Cooperation and competition in two-person experimental games. Like the Kitty Genovese case, many contemporary real-world problems reflect situations in which individuals must make decisions under conflict between their own interests and that of others. Much of the use of analogues in social research involves experimental games designed to study when and under what conditions individuals choose to cooperate with another individual when their interests are in potential conflict.⁴

One commonly used experimental game is based on the “prisoner’s dilemma” structure, which receives its name from the following scenario: Two suspects are taken into custody and separated while being questioned. The district attorney is certain they are guilty of a specific crime but does not have adequate evidence to convict them at a trial. He points out to each prisoner that each has two alternatives: to confess to the crime the police are sure they have committed, or not confess. If they both do not confess, then the district attorney states he will book them on some very minor trumped-up charge, such as petty larceny and illegal possession of a weapon, and they will both receive minor punishment; if they both confess, they will be prosecuted but he will recommend less than the most severe sentence. If one confesses and the other does not, then the confessor will receive lenient treatment for turning state’s evidence, whereas the latter will get the book slapped at him (from Luce & Raiffa, 1958).

In this scenario, the maximal joint outcome for both prisoners is attained if neither confesses. However, each of the individuals has to face the risk that if one refuses to confess while the other does confess, the refuser’s outcome will be very bad. Thus, each is motivated to attempt to maximize personal gain by confessing. If both individuals act on this motivation (as the district attorney wants and expects them to do), their joint outcome will be less than optimal. The prisoner’s dilemma is represented in the laboratory in the form of a “nonzero sum” or “mixed-motive” two-person game. On each trial of such a game, each of the two players makes a choice between two options and the payoff or winnings from the choice is determined by the choice made (simultaneously) by the other player. The potential choices and outcomes are represented by the joint payoff matrix, shown in Figure 6.3, which is one of many examples of payoff matrices that could be constructed in developing prisoners’ dilemma games.

Depending on the joint choices of the two participants, the first value in each cell represents player A’s payoff or winnings (in coins or chips) and the second represents player B’s winnings. In this type of matrix, choice 2 is the selfish and self-serving choice for each player, if completely ignoring the other player’s choice. However, if both players are selfish and decide on choice 2, their joint payoff will be only one coin each. The maximum joint payoff of three coins for each player can be achieved only when both players choose choice 1, but if one player chooses choice

		Player B's choice	
		1	2
Player A's Choice	1	3, 3	0, 5
	2	5, 0	1, 1

FIGURE 6.3 Prisoner’s dilemma game matrix.

1 while the other sticks with choice 2, the resulting payoffs will be highly uneven. Thus, the optimization of income across both players collectively can be achieved only through cooperation in which each player can choose choice 1 with some degree of confidence that the other player will do so too. In the typical prisoner's dilemma study, the dependent measure is the total number of competitive (choice 2) and cooperative (choice 1) choices made by each player, along with the measure of total joint payoffs achieved by each dyad.

Analogue experiments of collective decisions. Conflicts between individual interests can also take place at the group level. Dwindling fuel supplies, electrical brownouts, depletion of fresh water supplies, and air pollution are all cases in which individuals acting in their own self-interest can lead to collective disaster. One solution to the negative consequences of such social dilemmas requires that individuals restrain their own self-interested behavior (take shorter showers, carpool to work, reduce heat or cooling in their homes) in the interest of the collective good. Unilateral restraint on the part of any single individual is useless, however, unless a large proportion of other individuals exercise the same restraint. Yet, if many others exercise such constraint, a single individual doesn't need to and, hence, the dilemma.

Because it is difficult to design an experiment that mirrors the features of large-scale social problems such as the conservation of energy and other scarce resources, various stripped-down versions of social dilemmas have been designed for the laboratory to study individual and group decision-making processes (see Balliet et al., 2011; Messick & Brewer, 1983). In a sense, the Prisoner's Dilemma Game (PDG) described earlier is a two-person version of such collective choice analogues. However, research using the PDG came under heavy criticism for losing sight of any parallels between decision-making in the PDG setting and real-world decision-making contexts. To avoid falling into the same trap, analogue experiments of collective decision problems have been designed with careful attention to their real-world counterparts.

Of the collective dilemmas that have been designed for laboratory investigations, the most clearly analogous to conservation situations is the replenishable resource task, versions of which have been used in several group studies (Allison & Messick, 1985; Kramer & Brewer, 1984; Maldonado & Moreno-Sanchez, 2016; Messick et al., 1983; Milinski et al., 2008). The basic structure of this task involves the existence of a common pool of points (worth something of value such as money or experimental credits) to which each of the participants in an experimental session has access. On each trial or round of the task, every participant is permitted to draw from the common resource pool a selected number of points (which becomes theirs to keep) up to some maximum limit. After each round, the pool is replenished by some proportion of the pool size remaining after participants have drawn off their portions. The replenishment rate is set in such a way that *if* the participants as a group restrain their total take on each trial to a level below the maximum possible, the resource pool size can be maintained at its original level indefinitely. However, if the total take on each trial exceeds the rate of replenishment, the pool will be depleted gradually until the common resource is completely exhausted.

The participants in this resource dilemma situation are not being asked to *act as if* they were making decisions about conservation of energy or some other simulated experience. Rather, they are making actual decisions about real scarce resources in the scaled-down laboratory setting. Researchers hope that the task has been structured so that the basic elements of the decision to be made are the same as those operative in the dilemma situations that exist outside the lab. As in real large-scale collective problems, participants must make their individual decisions about resource use in the absence of knowledge about what other participants are doing and without coordination of choices. If these structural elements accurately parallel those in the real-world situation, then researchers can use the laboratory analogue to determine how variations in contextual features (e.g., availability of feedback about the state of the resource pool,

forms of communication among participants, identity of the group members, etc.) can alter the decisions that are made on the collective level. Such research can be used simultaneously to test hypotheses about basic psychological or group processes and about the effectiveness of potential interventions that might be implemented in the real world to influence the conservation of scarce resources (Messick & Brewer, 1983; Torres & Macedo, 2000).

Analogue experiments also have been designed to mimic other real-world social issues such as intergroup cooperation (Gaertner et al., 1989), school desegregation (Brewer & Miller, 1984), and bribery (Cameron et al., 2009). Even with several structural embellishments, it should be noted that the laboratory analogue experiment will always be a “stripped-down” representation of the corresponding situations in the real world. It is practically impossible to capture the historical and cultural context of social issues in society at large and fully translate these issues into an analogue experiment. Nonetheless, experience with the analogue paradigm indicates that it can engage – at least temporarily – much of the emotional significance attached to real decision-making. Given this level of involvement, the analogue experiment can provide a low-cost method for testing the potential effectiveness of various intervention strategies designed to reduce social problems such as intergroup discrimination, destruction of the environment, or organizational corruption. By utilizing analogue experiments in this way, promising avenues of policy research can be identified, and potentially costly mistakes avoided in the social arena.

Conclusion

Laboratory experiments can take many forms, but their basic requirements are similar across different types of studies. At the most basic level, a participant pool must be selected, the sample size must be decided upon, and materials must be prepared and vetted by the appropriate ethics review committee. The type and combination of treatments or manipulations to be used, environmental, stimulus, social, instructional, or prime, will depend upon the issue under investigation. In planning any experiment, issues of mundane and experimental realism must be considered carefully, and treatments should be designed to maximize both of these essential features.

Sometimes, the laboratory experiment can take the form of a role-playing simulation, in which the participants adopt different “roles” that might otherwise not be easily or ethically studied. Role-playing research can involve active or passive role players and can provide important insights into factors that are otherwise difficult to manipulate. Analogue studies are also a common research form, in which specific features of a real-world social context are represented by analogous features in the experimental setting. In this sense, they are not role-playing, but naturally responding to a stripped-down parallel of the social environment, allowing the researcher to study very specific responses to equally specifically controlled stimuli. These approaches seek to extend the reach of the laboratory study and are designed to lend elements of realism to participants’ behaviors.

Notes

- 1 Following recommendations by Cohen (1988), differences between two groups of 0.20, 0.50, and 0.80 standard deviation units are defined as small, moderate, and large effect sizes, respectively.
- 2 Earlier research had shown that respondents thought Republican candidates were more likely to brandish the American flag, but participants did not believe this would affect their voting behavior. Carter’s research suggests the respondents were incorrect.
- 3 In some types of simulation research, the human participant is replaced by a computer model of the processes under investigation. Such computer simulations are very valuable tools for theory

development (cf. Abelson, 1968; Hastie & Stasser, 2000). Their status as empirical research, however, is ambiguous and such forms of simulation are not covered here.

- 4 A whole field of “experimental economics” (or behavior economics) has developed around these types of experimental games and game theory (Durlauf & Blume, 2010).

References

- Abelson, R. P. (1968). Simulation of social behavior. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (2nd ed., vol. 2, pp. X–XX). Reading, MA: Addison-Wesley.
- Adobor, H., & Daneshfar, A. (2006). Management simulations: Determining their effectiveness. *Journal of Management Development*, 25, 151–168.
- Allison, S. T., & Messick, D. M. (1985). Effects of experience on performance in a replenishable resource trap. *Journal of Personality and Social Psychology*, 49, 943–948.
- Anderson, C. A. (1989). Temperature and aggression: Ubiquitous effects of heat on occurrence of human violence. *Psychological Bulletin*, 106, 74–96.
- Anderson, C. A. (2001). Heat and violence. *Current Directions in Psychological Science*, 10, 33–38.
- Aronson, E., & Carlsmith, J. M. (1968). Experimentation in social psychology. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (2nd ed., Vol. 2, pp. 1–79). Research methods. Reading, MA: Addison-Wesley.
- Asch, S. E. (1948). The doctrine of suggestion, prestige, and imitation in social psychology. *Psychological Review*, 55, 250–277.
- Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgment. In H. Guetzkow (Ed.), *Groups, leadership, and men* (pp. 222–236). Pittsburgh: Carnegie Press.
- Balliet, D., Mulder, L. B., & Van Lange, P. A. M. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin*, 137(4), 594–615.
- Bargh, J. A., & Chartrand, T. L. (2000). Studying the mind in the middle: A practical guide to priming and automaticity research. In H. Reis & C. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 253–285). New York: Cambridge University Press.
- Bem, D. J. (1965). An experimental analysis of self-persuasion. *Journal of Experimental Social Psychology*, 1, 199–218.
- Bem, D. J. (1967). Self-Perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review*, 74, 183–200.
- Berkowitz, L., & Lepage, A. (1967). Weapons as aggression-eliciting stimuli. *Journal of Personality and Social Psychology*, 7(2, Pt.1), 202–207.
- Breau, D. L., & Brook, B. (2007). ‘Mock’ mock juries: A field experiment on the ecological validity of jury simulations. *Law & Psychology Review*, 31, 77–92.
- Brehm, J. W., & Cohen, A. R. (1959). Re-evaluation of choice-alternatives as a function of their number and qualitative similarity. *Journal of Abnormal and Social Psychology*, 58, 373–378.
- Brehm, J. W., & Cohen, A. R. (1962). *Explorations in cognitive dissonance*. New York: Wiley.
- Brewer, M. B., & Gardner, W. (1996). Who is this “we”? Levels of collective identity and self-representations. *Journal of Personality and Social Psychology*, 71, 83–93.
- Brewer, M. B., Manzi, J. M., & Shaw, J. S. (1993). In-group identification as a function of depersonalization, distinctiveness, and status. *Psychological Science*, 4, 88–92.
- Brewer, M. B., & Miller, N. (1984). Beyond the contact hypothesis: Theoretical perspectives on desegregation. In N. Miller & M. B. Brewer (Eds.), *Groups in contact: The psychology of desegregation*. Orlando, FL: Academic Press.
- Cameron, L., Chaudhuri, A., Erkal, N., & Gangadharan, L. (2009). Propensities to engage in and punish corrupt behavior: Experimental evidence from Australia, India, Indonesia, and Singapore. *Journal of Public Economics*, 93, 843–851.
- Carlson, R. (1971). Where is the person in personality research? *Psychological Bulletin*, 75, 203–219.
- Carter, T. J., Ferguson, M. J., & Hassin, R. R. (2011). A single exposure to the American flag shifts support toward Republicanism up to 8 months later. *Psychological Science*, 22, 1011–1018.
- Cohen, J. A. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cohen, J. A. (1992). A power primer. *Psychological Bulletin*, 112, 155–159.

- Correll, D. N., Engle, K. M., Lin, S. S., Lac, A., & Samuelson, K. W. (2021). The effects of military status and gender on public stigma toward posttraumatic stress disorder. *Stigma and Health*, 6(2), 134–142.
- Cox, J., Clark, J. C., Edens, J. F., Smith, S. T., & Magyar, M. S. (2013). Jury panel member perceptions of interpersonal-affective traits of psychopathy predict support for execution in a capital murder trial simulation. *Behavioral Sciences & the Law*, 31(4), 411–428.
- Crano, W. D., & Prislín, R. (2006). Attitudes and persuasion. *Annual Review of Psychology*, 57, 345–374.
- Crano, W. D., Siegel, J. T., Alvaro, E. M., & Patel, N. (2007). Overcoming adolescents' resistance to anti-inhalant appeals. *Psychology of Addictive Behaviors*, 21, 516–524.
- Cronkite, R. C. (1980). Social psychological simulations: An alternative to experiments? *Social Psychology Quarterly*, 43, 199–216.
- Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology*, 10, 202–214.
- Durlauf, S. N., & Blume, L. E. (Eds.) (2010). *Behavioural and experimental economics*. London: Palgrave Macmillan.
- Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *Journal of Personality and Social Psychology*, 58, 203–210.
- Fowler, S. M., & Pusch, M. D. (2010). Intercultural simulation games: A review (of the United States and beyond). *Simulation & Gaming*, 41, 94–115.
- Freedman, J. L. (1969). Role playing: Psychology by consensus. *Journal of Personality and Social Psychology*, 13, 107–114.
- Gaertner, S. L., Mann, J., Murrell, A., & Dovidio, J. (1989). Reducing intergroup bias: The benefits of recategorization. *Journal of Personality and Social Psychology*, 57, 239–249.
- Galtier, F., Bousquet, F., Antona, M., & Bommel, P. (2012). Markets as communication systems: Simulating and assessing the performance of market networks. *Journal of Evolutionary Economics*, 22, 161–201.
- Gamson, W. A. (1969). *SIMSOC, simulated society: Participant's manual with selected readings*. New York: MACM.
- Gamson, W. A., & Stambaugh, R. J. (1978). The model underlying SIMSOC. *Simulation & Games*, 9, 131–157.
- Gardner, W. L., Pickett, C. L., & Brewer, M. B. (2000). Social exclusion and selective memory: How the need to belong influences memory for social events. *Personality and Social Psychology Bulletin*, 26, 486–496.
- Gerard, H. B., Wilhelmy, R. A., & Conolley, E. S. (1968). Conformity and group size. *Journal of Personality and Social Psychology*, 8, 79–82.
- Greenwood, J. D. (1983). Role-playing as an experimental strategy in social psychology. *European Journal of Social Psychology*, 13, 235–254.
- Hamilton, D. L., Katz, L. B., & Leirer, V. O. (1980). Cognitive representation of personality impressions: Organizational processes in first impression formation. *Journal of Personality and Social Psychology*, 39, 1050–1063.
- Hastie, R., & Stasser, G. (2000). Computer simulation methods for social psychology. In H. Reis & C. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 85–114). New York: Cambridge University Press.
- Hofstede, G. J., & Murff, E. J. T. (2012). Repurposing an old game for an international world. *Simulation & Gaming*, 43, 34–50.
- Horowitz, I. L., & Rothschild, B. H. (1970). Conformity as a function of deception and role playing. *Journal of Personality and Social Psychology*, 14, 224–226.
- Janis, I. L., & Mann, L. (1965). Effectiveness of emotional role-playing in modifying smoking habits and attitudes. *Journal of Experimental Research in Personality*, 1, 84–90.
- Kelman, H. C. (1967). Human use of human subjects: The problem of deception in social psychological experiments. *Psychological Bulletin*, 67, 1–11.
- Klein, R. D., & Fleck, R. A. (1990). International business simulation/gaming: An assessment and review. *Simulation and Gaming*, 21, 147–165.
- Kramer, R. M., & Brewer, M. B. (1984). Effects of group identity on resource use in a simulated commons dilemma. *Journal of Personality and Social Psychology*, 46, 1044–1057.
- Lac, A., & Donaldson, C. D. (2021). Experimental priming of peer injunctive norms and peer descriptive norms on personal alcohol attitudes, behaviors, and motivations. *Addiction Research & Theory*, 29, 338–346.

- Latané, B., & Darley, J. M. (1968). Group inhibition of bystander intervention. *Journal of Personality and Social Psychology*, 10, 215–221.
- Latané, B., & Darley, J. M. (1970). *The unresponsive bystander: Why doesn't he help?* New York: Appleton-Century Crofts.
- Lord, C. G., & Saenz, D. S. (1985). Memory deficits and memory surfeits: Differential cognitive consequences of tokenism for tokens and observers. *Journal of Personality and Social Psychology*, 49, 918–926.
- Luce, R. D., & Raiffa, N. (1958). *Games and decisions*. New York: Wiley.
- Maldonado, J. H., & Moreno-Sanchez, R. del P. (2016). Exacerbating the tragedy of the commons: Private inefficient outcomes and peer effect in experimental games with fishing communities. *PLoS One*, 11, 1–17.
- Messick, D. M., & Brewer, M. B. (1983). Solving social dilemmas: A review. In L. Wheeler & P. Shaver (Eds.), *Review of personality and social psychology* (Vol. 4). Beverly Hills, CA: Sage.
- Messick, D. M., Wilke, H., Brewer, M. B., Kramer, R. M., Zemke, P., & Lui, L. (1983). Individual adaptations and structural change as solutions to social dilemmas. *Journal of Personality and Social Psychology*, 44, 294–309.
- Milgram, S. (1963). Behavioral study of obedience. *The Journal of Abnormal and Social Psychology*, 67, 371–378.
- Milgram, S. (1965). Some conditions of obedience and disobedience to authority. *Human Relations*, 18, 57–75.
- Miller, A. G. (1972). Role playing: An alternative to deception? A review of the evidence. *American Psychologist*, 27, 623–636.
- Milinski, M., Sommerfeld, R. D., Krambeck, H.-J., Reed, F. A., & Marotzke, J. (2008). The collective-risk social dilemma and the prevention of simulated dangerous climate change. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, 105, 2291–2294.
- Mixon, D. (1972). Instead of deception. *Journal for the Theory of Social Behavior*, 2, 145–177.
- Niedenthal, P. M., Barsilou, L., Winkielman, P., Krauth-Gruber, S., & Ric, F. (2005). Embodiment in attitudes, social perception, and emotion. *Personality and Social Psychology Review*, 9, 184–211.
- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 231–259.
- Rosenberg, M. J., & Abelson, R. P. (1960). An analysis of cognitive balancing. In C. I. Hovland & M. J. Rosenberg (Eds.), *Attitude organization and change: An analysis of consistency among attitude components* (pp. 112–163). New Haven, CT: Yale University Press.
- Schultz, D. (1969). The human subject in psychological research. *Psychological Bulletin*, 72, 214–228.
- Tormala, Z. L., & Rucker, D. D. (2022). Attitude change and persuasion: Classic, metacognitive, and advocacy perspectives. In L. R. Kahle, T. M. Lowrey, & J. Huber (Eds.), *APA handbook of consumer psychology* (pp. 323–350). Washington, DC: American Psychological Association.
- Torres, M., & Macedo, J. (2000). Learning sustainable development with a new simulation game. *Simulation and Gaming*, 31, 119–126.
- Verplanken, B., & Orbell, S. (2022). Attitudes, habits, and behavior change. *Annual Review of Psychology*, 73, 327–352.
- Wells, G. L., & Petty, R. E. (1980). The effects of overt head movements on persuasion: Compatibility and incompatibility of responses. *Basic and Applied Social Psychology*, 1(3), 219–230.
- Xu, Y., & Bernard, A. (2013). A quantitative model on knowledge management for team cooperation. *Knowledge-Based Systems*, 45, 41–46.
- Zanna, M. P., & Cooper, J. (1974). Dissonance and the pill. An attribution approach to studying the arousal properties of dissonance. *Journal of Personality and Social Psychology*, 29, 703–709.
- Zimbardo, P. (2007). *The Lucifer effect: Understanding how good people turn evil*. New York: Random House.
- Zimbardo, P. G., Haney, C., Banks, W. C., & Jaffe, D. A. (April 8, 1973). Pirandellian prison: The mind is a formidable jailer. *New York Times Magazine*, 8, 38–60.

7

EXTERNAL VALIDITY OF LABORATORY EXPERIMENTS

It often is thought that the use of laboratory experiments in research involves achieving internal validity at the cost of external validity, or generalization of results to the world outside the laboratory. This view betrays a misunderstanding of the logic and meaning of internal and external validity, but it is common, and because of this, critics within and outside the social sciences have argued that the experimental approach that structures so many of our research endeavors is inadequate or inappropriate for the study of social beings in natural settings.

A commonplace criticism of the laboratory experiment concerns its artificiality, or reactivity, and the consequent impossibility of determining the adequacy of experiment-based generalizations. After reviewing nearly 20 years of research in his field, for example, Cronbach (1975, p. 116) appeared to have despaired even of the possibility of developing lasting and useful generalizations from social research when he stated,

It [is] unlikely that social scientists will be able to establish generalizations applicable beyond the laboratory or that generalizations established in the field work will be maintained. Research should be less concerned with hypothesis testing and more concerned with interpreting findings in local contexts.

A similar theme was sounded by Gergen (1973, 1976; but see Schlenker, 1974; Shadish et al., 2002; Wallach & Wallach, 1993), who argued that social science research was more a historical than a scientific enterprise. Our theories, he argued, are little more than post hoc descriptions restricted by the historical contexts in which they were developed. As circumstances changed, so too must these time-bound descriptions.

Campbell and Stanley (1963) claimed the generalizability of research findings was never settled definitively. We agree, but although the certainty of generalizations can never be attained, it can be approximated, and the approximations can become progressively more exact if we are sensitive to the forces that influence behavior within and outside the experimental laboratory. So, for the remainder of this chapter we will consider aspects of the laboratory experiment that affect the trustworthiness of experimentally based generalizations. Although our observations cannot answer all objections raised by the critics of experimentation in social research, the recommendations will provide firmer grounds for generalization if followed than if ignored. It

also will become clear that the issues of generalization raised in the response to experimental approaches apply also to the nonexperimental and quasi-experimental research tactics discussed in the upcoming chapters.

Generalizability across Participants

The principal ingredients of all laboratory experiments include participants, experimenters, measures, and manipulations employed to influence participants' behaviors. In the first part of this chapter, we concentrate on generalizability issues related to the participant sample and the experimental treatments to which participants are exposed. In this chapter's last part, we take a closer look at the actions of the experimenter and the potential impact of these behaviors on the internal as well as external validity of an experiment.

Restriction of Participant Populations

Many critics of social experimentation have argued that from the standpoint of generalization, the college undergraduate, typical of much research, is far from the ideal choice as the research sample. In fact, it has been stated that there is possibly nothing more dissimilar to the average "person in the street" than the college undergraduate. The average college student is more intelligent than the typical person on the street; usually healthier; more concerned with the social forces operating on the physical and social environment; sensitive to the various communication media; less likely to hold strongly fixed attitudes; and plays a more active role in exerting control over the factors influencing daily life (Sears, 1986). Furthermore, the college students that we do study are drawn almost exclusively from western, educated, industrialized, rich, and democratic societies (WEIRD: Henrich et al., 2010). However, even when college students are used as a primary source of information or as a comparator their responses do not differ, or differ only in minor ways, with groups more representative of the larger population (e.g., see Cemalcilar et al., 2021; Kjærviik & Bushman, 2021). In situations in which basic human processes are under study (and it is in precisely this type of situation that the experiment is most advantageously employed), it is reasonable to assume that the idiosyncrasies of the individuals who have been sampled randomly should have relatively little effect on experimental results. On the other hand, it is unreasonable to build a science on a single source of participants, whether they are college sophomores, men, or MTurk workers. Multiple indicators of a construct, multiple research contexts, and different types of participants help ensure results are generalizable and not dependent on a single operationalization, a constant research context, or a single type of participant.

Participant Awareness

Arguments about the representativeness of college student participants deal with only one aspect of the "participant" issue. Many critics of laboratory experimentation object not simply because a sample is drawn from an unusual population, but because the responses of any person conscious of being under observation could be quite different from those of persons not under such scrutiny. This argument is important, but the degree to which this "self-consciousness effect" might operate within any given experiment should be considered. In some cases, the participant's awareness of being observed may be the most salient feature of the total experimental situation. If this is so, then the obtained results of such a study should be viewed with caution – but certainly not dismissed from consideration. In more ideal instances, however, because the situation engages the attention of the participant, it seems likely that participants' awareness of

being studied would be greatly diminished, and thus would have negligible impact on research findings. In most cases, the experimenter should attempt to construct laboratory situations that reduce the self-consciousness of participants as much as possible.

One characteristic that can be expected to influence participants' awareness of being under study and how they respond to that awareness is freedom they had in agreeing to commit themselves to the research. Depending on how participants are recruited, they can be generally classified into one of three categories, which we have termed voluntary, involuntary, and nonvoluntary.

Voluntary participants. Individuals considered to be **voluntary participants** are aware they are under investigation but have made a conscious decision that the potential benefits outweigh the costs (measured in terms of time spent, privacy invaded, etc.) of being in the study. This decision can be prompted by many factors: monetary incentives, altruism, attainment of greater personal insight, or possibly contributing to the understanding of human behavior. If this positive mindset is maintained during the experiment, and the design and procedures do not impel participants to revise their estimates of benefits and costs, then it is doubtful that such participants would willfully attempt to subvert the experiment. This is not to suggest voluntary participants become unaware of the fact that they are being studied, rather that they consider this potential invasion of privacy to be part of the contract with the experimenter and are willing, as long as the cost/benefit ratio remains positive, to respond in an honest and authentic manner.

A more worrisome problem occurs when the voluntary participant proves too willing to cooperate, too eager to help the experimenter confirm research hypotheses. In an extensive review of the volunteer participant literature, Rosenthal and Rosnow (1975) found volunteers were better educated, had higher occupational status, higher need for approval, higher intelligence, and better adjustment than nonvolunteers. If these factors contribute to an overly cooperative volunteer, a participant intention to help confirm the research hypotheses could adversely affect a study's generalizability.

Involuntary participants. We define **involuntary participants** as research subjects who feel that they have been coerced to spend their time in an investigation and consider it unjustifiable, and therefore vent their displeasure by actively attempting to ruin the study. They have proven to be the bane of experiments. The feeling of coercion on the part of participants can be activated in many ways. In universities in the United States, for example, introductory psychology students often are expected to serve in some minimum number of research studies as part of course requirements. In earlier times in the armed forces, draftees were commonly tested, and sometimes experimented upon without their consent. Individuals of minority groups, because of their relative rarity, often find themselves the targets of unwanted scientific scrutiny. Persons forced to comply with the demands of some higher authority can generate considerable resentment, which can seriously affect the outcomes of studies. Dissatisfaction with "guinea pig" status is both understandable and justified. Unless the experimenter can demonstrate that participation in the study is not a one-sided proposition, but a cooperative venture in which both parties can gain something of value, there is no reason to expect willing acquiescence to the participant role. Sometimes payment can be used to help the participant justify participation in a study. More satisfactory is the experimenter's explanation of the reasons for, and the importance of the investigation prior to participation in the experiment. Participants' willingness to serve in studies they perceive as important is gratifying and sometimes surprising.

Whether participants are more accurately described as voluntary or involuntary, some rationale for their participation always should be provided at the beginning of the experiment. No one likes to waste time, even if paid to do so. By informing participants of the importance

of the research project and of the importance of their role in it, the experimenter can instill a sense of positive commitment in participants. In most circumstances, but not always, informing participants of their role is a prerequisite for their participation. Most ethics review committees require at the beginning of the study that participants be apprised of their rights, which includes information regarding the time involved, the potential risks their work might entail, and the potential contribution of their efforts.

Nonvoluntary participants. Those considered **nonvoluntary participants** unknowingly participate in a study and are unaware of their role in the research. Often, they are not informed of their “participation” until the completion of the study – and sometimes they never learn that their behaviors constituted the raw material of a scientific investigation. After their responses are recorded, the investigator may (or may not) explain the nature of the study to the “participants.” Because they are unaware of their participation in a study, it is obvious that the reactions of nonvoluntary participants cannot be described as artificial or laboratory dependent. Results based on the reactions of such individuals should enjoy a high degree of generalizability. Many issues can be investigated experimentally in naturalistic surroundings with nonvoluntary participants, if one is attentive to the investigative possibilities offered by contemporary events. For example, Sivacek and Crano (1982) polled students on several issues of concern, including a proposed referendum to raise the legal drinking age. Two weeks later, the same respondents were re-contacted and asked if they wished to work for a (fictitious) organization to help defeat the referendum. At this point, the participants did not realize they were still a part of a psychological study. As expected, even though students of all ages had the same negative attitude toward the law change, younger participants, who were most vested in the issue, were significantly more willing to pledge more of their time than older respondents, for whom the law’s change would have little impact.

Nonvoluntary participants are typically used in field research situations, outside the laboratory, where they are not made aware that they are taking part in an experiment. However, even within the context of laboratory experiments, some studies have been designed so that the participants entering (or leaving) an experimental laboratory are treated and tested before they realize the experiment has begun (or after they think it has ended). Bargh and associates (1996) provided an interesting example of the use of nonvoluntary participants in their intriguing study of the effects of priming (which is discussed in Chapter 18). In their experiment, they activated the stereotype of the elderly to determine its effects on subsequent behavior in a sample of young college students. Researchers gave the participants 30 sets of five words. In an apparent test of language proficiency, they were instructed to rearrange and unscramble each set of words so they made grammatical sense. In the “elderly prime” condition, the sentences contained words associated with old age, such as *old*, *lonely*, *gray*, *forgetful*, *retired*, *wrinkle*, *ancient*, *Florida*, and *bingo*, among others. For the control group, these words were replaced by words that were unrelated to age (thirsty, clean, etc.).

After participants individually completed the task, they were partially debriefed and dismissed. They were told the elevator to leave the building was down the hall. When they left the office, a confederate activated a stopwatch and timed their journey from the doorway to the carpet in front of the elevator. At this point, the experimenter intercepted the participant and fully explained the study. The results showed that participants who had been randomly assigned to the elderly prime took significantly longer to walk to the elevator (as would an elderly person) than those in the control group.

This study provides an interesting examination and extension of the nonvoluntary participant concept. Obviously, participants in Barge and associates’ experiment knew they were the subjects in a psychological experiment. However, the critical behavior was assessed after they

thought the experiment was completed, so it is unlikely that their speed of walking to the elevator was affected by their awareness of being the subject of the earlier psychological study.

Should the behavior of nonvoluntary “participants” ever be studied? Clearly, these individuals have not been provided the opportunity to provide their informed consent to engage as participants in a study. The rationale for justifying the use of nonvoluntary participants is that their behavior if emitted in public implies consent. This argument rules out behaviors where privacy is assumed. We discuss the ethical implications of this issue in Chapter 21.

Participant Roles

Related to the voluntary-involuntary-nonvoluntary participant distinction, participants can also be categorized in terms of the way they perceive their role in the research study and how they enact that role. For example, Webber and Cook (1972) proposed and tested four participant types: *good participants*, who attempt to determine the experimenter’s hypotheses and attempt to confirm them; *negative participants*, who also are interested in learning the study hypotheses so they can disconfirm them; *faithful participants*, who are willing to cooperate with almost any of the experimenter’s demands and follow instructions scrupulously, ignoring any hypotheses they might have formed about study expectations; and *apprehensive participants*, who worry the experimenter will use their performance to evaluate their intelligence, morality, etc., and respond to experimental demands accordingly.

It is reasonable to assume almost all participants are at least partly apprehensive about taking part in a psychological study, although this apprehension probably diminishes with experience. Involuntary or negative participants are most likely to be antagonistic, as are voluntary participants who learn that the cost to benefit ratio of their participation is not as favorable as they thought originally. The good and the faithful participant roles are most likely to be assumed by voluntary participants. Nonvoluntary participants, being unaware that they are being studied, are unlikely to assume any of the roles Webber and Cook (1980, p. 507) defined. Research on participant perceptions (e.g., Carpio et al., 1983; Strohmets, 2008) has produced evidence supportive of Webber and Cook’s categorization model, particularly for the “good” and “faithful” participant roles. For example, studies of participants’ motivation in experiments have found that many would go to great lengths to “please” the experimenter to help confirm perceived research hypotheses. Although Orne (1962) was not the first to investigate this issue (e.g., see Pierce, 1908), he was able to demonstrate quite clearly the almost incredible degree to which participants were willing to foster his research expectations, as they perceived them. Orne presented participants with approximately 2,000 *pages* of random numbers, instructed them to sum adjacent numbers – a total of 224 additions performed on each page – and to continue to work at this task until his return. Incredibly, Orne found almost no one willing to quit this task even after 5 *hours* had elapsed. Adding a final step to this process – asking participants to tear the completed pages into “a minimum of 32 pieces,” throw the pieces into the waste basket, and begin again – had almost no effect on participants’ persistence. In other studies, participants continued other similarly self-defeating tasks beyond reasonable limits. When questioned about their unusual perseverance, they often responded with a guess that the research hypothesis was concerned with endurance, and thus their actions were quite appropriate. Performances of this type have alerted researchers to the tremendous degree of behavioral control the experimenter can intentionally or unintentionally exert in the laboratory.

Nowhere was this potential more evident than in the studies of Stanley Milgram (1963, 1965). While investigating the effects of a range of variables on “obedience,” Milgram’s basic experimental procedure took the guise of a two-person (teacher-learner) verbal memory and

learning study. In these studies, the naïve participant was asked to serve as the “teacher,” whose task was to shock the “learner” (confederate) each time the learner committed a recall error from a list of words. In fact, a confederate played the role of the learner, and received no shocks.

The shock generator was a large, formidable apparatus with 30 switches and voltage levels in 15-volt increments ranging from 15 to 450 volts, and written descriptions of available voltages ranging from “Slight Shock” to “Danger: Severe Shock.” The confederate was placed in an adjacent room and purposely gave many wrong answers based on a predetermined script to standardize the study across participants. In response to each incorrect answer, the teacher (participant) was to increase the level of punishment.

The dependent measure of this study was the voltage level at which the naïve participant refused to administer further punishment to the “learner.” When the 150-volt level was reached, the learner audibly demanded to be set free, to terminate the experiment. Although the learner was unseen, his protestations, amplified from the adjacent room, were truly heartrending. After this point in the investigation, almost all the participants had requested to end the study. The researcher, however, always responded, “You have no other choice, you must go on!” A substantial percentage of the participants obeyed all the researcher’s commands. In one of these studies, for example, 61% of the participants continued to the very end of the shock series (Milgram, 1965); and across the entire study series, the mean compliance rate was in the 60% range!

When questioned about their actions afterward, Milgram’s participants’ responses were like those of Orne’s. The most common rationale offered by participants was that they viewed themselves as an integral part of a scientific investigation, and they were determined to fulfill their commitment to the study, even though these participants did not seem to enjoy themselves in this study. Milgram’s filmed records indicate that many naïve participants vehemently protested the demands to continue the experiment, and demonstrated clear signs of nervousness, stress, and tension. Even so, most continued to the bitter end of the shock generator. Under the auspices of a “scientific investigation” people seemed willing to obey the commands of the authority figure and commit themselves to boring, tedious, meaningless, and even blatantly reprehensible actions, which they well might have refused to perform outside the laboratory setting.

Apart from the moral and ethical implications of this investigation, we also must consider findings of this type from the standpoint of generalizability. Were the participants willing to perform their questionable behaviors only when asked to do so for the sake of science? If so, we must question whether the relationships discovered would hold outside the laboratory setting. But suppose the participants *were* willing to assume these roles outside the laboratory and perform as told whenever an important authority (duty, allegiance to a higher law, etc.) was involved. In that case, Milgram’s research would be relevant to a variety of contexts.

Experimenter Expectancy and Bias

To this point, we have shown how decisions regarding the basic features of the experiment can influence its internal and external validity. Should we use voluntary, involuntary, or nonvoluntary participants? Should we be concerned with experimental realism or mundane realism (discussed in the previous chapter)? Recall that *mundane realism* refers to the extent to which the research setting and operations resemble events in normal, everyday life. *Experimental realism* is concerned with the impact of the experimental treatment. Is the experimental manipulation powerful enough that participants notice it and feel involved in the action of the experiment? These factors, generally under the control of the experimenter, have great bearing on the quality and ultimate contribution of the research enterprise. However, in addition to these deliberate choices, there are aspects of the setting that are generally not consciously controlled, but that can

have considerable influence on the quality and generalizability of a study's results. These influences originate in the experimenter or the experimental setting. For example, we have learned that the mere presence of the experimenter can operate as a subtle but nevertheless potentially powerful "treatment," affecting participants' responses as a result not of the experimental manipulation, but of the experimenter's own expectations about the study.

In an important series of studies initiated in the 1960s, Robert Rosenthal and associates demonstrated that the expectations held by a researcher seemingly can be transmitted to participants, be they college sophomores (Rosenthal & Fode, 1963) or elementary school children (Rosenthal & Jacobson, 1968). Over the years, Rosenthal continued his investigations of experimenter expectancy effects by showing their effects on members of juries (Rosenthal, 2002a), geriatric patients (Ambady et al., 2002), counseling patients (Rosenthal, 2002b), even Sprague-Dawley albino rats (Rosenthal & Lawson, 1964), among others. Although controversy still surrounds the medium of communication involved in producing such effects, the effect has been demonstrated over a range of experimental and correlational contexts (Harris & Rosenthal, 1985, 1986; Rosenthal, 1966, 2002a, 2002b, 2003).

Although a major contribution to contemporary psychological research, Rosenthal's research on expectancy effects was not new, as evident in studies of Moll (1898) and Jastro (1900) of the early twentieth century, who devoted serious attention to the issue. In fact, the "double-blind" method in pharmacological research (in which neither patient nor physician know whether a real drug or a placebo has been administered) was developed specifically to combat expectancy effects on the part of patients and physicians. A **placebo effect** occurs when participants' belief in the efficacy of the treatment, rather than the actual effects of treatment, is responsible for the results found in a treatment or experiment. Experimenters test placebo effects by administering an inert "treatment" superficially resembling the real one, to keep participants ignorant of whether they received the treatment or not. In pharmacological research, participants in the placebo group typically ingest an inert pill that physically resembles the actual drug. If the experimental and placebo groups exhibit the same outcome scores, this result suggests that the actual treatment itself is not effective, at least not beyond that motivated by participants' beliefs about the treatment's effectiveness.

In an early investigation of placebo effects, Beecher (1966) compared the effects of morphine, a powerful narcotic, with those of a placebo in the management of moderate pain. Using a double-blind procedure, no differences in pain alleviation were detected between the experimental and control (placebo) groups. Similarly, Reed and Witt (1965) apparently were able to induce hallucinations on the part of a participant who thought he had been given the hallucinogenic drug LSD when, in fact, a placebo had been administered. De la Fuente and colleagues (2001) found strong placebo effects on dopamine release in patients with Parkinson's disease (PD): "The findings indicate that the **placebo effect** in PD is powerful and is mediated through activation of the damaged ... dopamine system" (p. 387). If effects this profound can be produced through manipulations of participants' treatment expectations, it is reasonable to argue that unintended, inadvertent behaviors on the part of experimenters may have similarly powerful effects on the outcomes of the research.

Rosenthal's early research convincingly demonstrated the presence of experimenter expectancy effects in the social research laboratory (Rosenthal and Fode, 1963). Using a group of students in his experimental psychology course, Rosenthal conducted a laboratory exercise in which selected students were to serve as experimenters. Each experimenter's task was to present a set of ten full face photographs to volunteer participants from an introductory psychology course. The participants were asked to rate the "successfulness" of the people depicted in the photographs. Up to this point, the study is completely straightforward. However, before they

began, Rosenthal told a cover story to half his experimenters, indicating that previous research had demonstrated that the photos depicted successful people, and thus, they could expect relatively high ratings from participants. The other experimenters administered the same photos, but were told the opposite. [In fact, the photos were from a larger pool of pictures chosen because they had been previously rated as neutral with respect to “successfulness.”] All experimenters then were told that because the lab exercise was designed to give them practice in “duplicating experimental results,” they would be paid twice the usual amount (\$2/hour, rather than \$1) if their results agreed with these previous findings.

The ratings the two groups of experimenters obtained from their participants were substantially different, in ways consistent with the expectancy hypothesis. Experimenters led to expect positive ratings from the participants obtained significantly more positive ratings than experimenters who expected negative scores. In a variation of this study, in which the biased experimenters were removed to the extent that they did not even handle the photographs and were not in direct facial contact with the participants, Fode (1960) obtained results like those of the first investigation. Adair and Epstein (1968) removed the biased experimenter even farther from the experimental situation by recording the instructions read by experimenters who had been led to expect either high or low ratings. They played these instructions to a new sample of participants and obtained expectancy effects in both the face-to-face and the tape-recorded instruction conditions. This result indicates that even unintentional vocal intonation and cues of experimenters may inform participants about the hypothesized and desired direction of results.

As experimenters were offered a monetary incentive to produce findings consistent with “previous research,” the question remains as to whether experimenters deliberately misrecorded responses in the confirmatory direction. Rosenthal has argued that although intentional recording errors on the part of the student experimenters were possible in these studies, they were unlikely to occur. Because the setting allowed participants to observe the experimenter recording their ratings, Rosenthal reasoned that any errors would have been detected and corrected immediately by the participants. Later studies in which the same general approach was employed have resulted in generally confirmatory findings of robust expectancy effects (e.g., Harris, 1991; Harris & Rosenthal, 1985; Rosenthal & Rubin, 1978).

Although the effects of *experimenter bias* due to recording, observation, and computation errors are troublesome, these effects can be detected and avoided (Simonsohn, 2013). More subtle and troublesome is the possibility in which a participant, through subtle and unintentional cues from the experimenter, decides to perform “correctly,” that is, in a manner that the participant thinks will please the researcher. The interactive nature of the experimenter/participant relationship renders control extremely difficult in such situations. Suppose each time a “correct” response was made, the experimenter reinforced the participant (verbally or nonverbally). The reinforcement could be extremely subtle and unintentional. A slight nod of the head, a fleeting smile, an almost imperceptible lean of the body toward the participant might be all that is needed to trigger a cooperative response. Returning to Rosenthal and Fode’s (1963) photo rating task, consider a participant actively attempting to “psych-out” the study when realizing that the experimenter gives an almost imperceptible smile (a micromomentary expression) each time a high photo rating (or a low one, depending upon the experimental condition assigned) is given.

Experimental demand. Cues about the researcher’s expectations are not limited to ones emitted directly by the experimenter but could involve artifacts in a laboratory setting that are not essential to the experiment, but that inadvertently provoke some form of participant response. For example, a “Just Say No to Drugs” poster carelessly posted on the doorway of a laboratory that investigates participants’ attitudes and behaviors toward use of illicit substances

may affect responses to items querying their drug use. An undergraduate research assistant wearing a LA Dodgers baseball hat may stimulate responses (positive or negative) irrelevant to the task, which may involve altruistic behavior toward the homeless. These so-called **demand characteristics** are the totality of all social cues communicated in an experimental context (laboratory or field), which are not a part of the intended manipulation, including those emanating from the experimenter and the setting, and that may alter and therefore place a demand on (i.e., influence) participants' responses. When participants are influenced by such demands, which may explicitly or implicitly suggest the preferred response, the study is said to suffer from demand characteristics. Many expectancy findings can be explained in terms of the behaviors of an experimenter too eager to obtain evidence that will confirm expectations and hypotheses and careless staging of the experimental setting, in combination with research participants who are too eager to please (or hassle) the experimenter. Pseudo-confirmations (or disconfirmations) of expectations can result from experimenters who:

- Systematically err in observation, recording, or analysis of data (whether intentionally or not),
- Cue the participant to the correct response through some form of verbal or nonverbal reinforcement, or
- Are not sufficiently thoughtful in staging the research context (lab or field) in such a way that unwanted, inappropriate cues to behavior are not present.

Solutions to the Problems of Experimenter Expectancy Bias

The design problem that must be overcome in controlling demand effects is twofold. It consists of controlling both experimenter bias and the intentional (or unintentional) cuing of the experimental participant. By properly controlling one of these factors, we sometimes control the other.

Monitoring. From the earlier discussion of observer bias, an important control entails careful observation of experimenters' behaviors to ensure that their data transcription and analysis are accurate. This may be accomplished by recording the experimenter-participant interaction and comparing the results found by the experimenter with those obtained through the unbiased observers of this interaction. Unfortunately, this process does not preclude the possibility of subtle, nonverbal cuing of the participant by the experimenter. Although a more rigorous observation of the social exchange between experimenter and participant is certainly worthwhile, it does not solve the potential expectancy problem.

Blind procedures. A more effective solution is the use of **double-blind procedures**, in which participants as well as the researcher responsible for administering the study are made unaware of the treatment condition to which participants are (randomly) assigned. This approach is used to control for the effects of both researcher bias and participant cuing, and borrows heavily from pharmacological research. The most obvious way this control can be achieved is through the simple expedient of not informing the experimenters who interact directly with participants about the aims of the research hypotheses or the condition to which the participants were assigned. This control is useful, and widely used, because if experimenters do not know what is expected, they will be unlikely to pass on any biasing cues to participants.¹ Similarly, any recording or calculation errors that might be made should be unbiased, and thus would not systematically influence results.

Unfortunately, experimenters, even naïve hired experimenters specifically shielded from information about the theory under development, have been shown to be hypothesis-forming

beings who, over the course of an investigation, might evolve their own implicit theory of the meaning and nature of the work they are performing. This possibility can be assessed through an investigation of the variability or spread of scores of the experimental data over the course of the study. If in a particular condition, scores tend to become more and more homogeneous as the investigation progresses, an implicit hypothesis may have been formed by the experimenter as the study progressed (with the attendant expectancy problems). Although such an analysis enables us to evaluate the possibility of the occurrence of implicit hypothesis formation, it does not suggest any realistic way of correcting for this effect.² Thus, the application of the experimental blind procedure would not seem to offer a complete solution to the experimenter expectancy problem, unless the entire experiment could be completed before a researcher could realistically develop a series of implicit hypotheses. It is likely that this hypothesis generation could be impeded if the experimenter was prohibited from discovering the hypothesized direction of results until the completion of the investigation, but often because of information conveyed by the materials and measures used in an experimental situation, this is not possible.

A slight variant of this procedure does not presume to eliminate knowledge of the research hypotheses from the investigators but limits their information about the research condition to which any participant or group of participants has been assigned. Often, in a variant of the double-blind experiment, it is possible to test both treatment and control participants in the same setting, at the same time, without the experimenter's or participants' knowledge of the specific condition into which any individual falls. This is especially likely when written or video instructional manipulations, delivered individually via lab computers, are used. For example, suppose that a market researcher wanted to decide between four spokespersons for a newspaper advertising campaign designed to help sell a new automobile model. The researcher might develop four newspaper advertisements with an identical message designed to sell the new car, except that they are attributed to different sources: a noted consumer advocate, a Nobel Prize-winning physicist, a trusted TV personality, and a popular sports hero. To avoid problems of experimenter expectancy, the market researcher randomly mixes the four newspaper ads, and distributes them to participants whose reactions to the new car model are assessed. Each participant in the room is presented an envelope containing the advertisement corresponding to one of the four conditions. Because the researcher does not know which participant received which source, it is obvious that communicated experimenter expectancies could not influence the results. Thus, even though the researcher might hold a strong expectation that the consumer advocate's endorsement would prove most persuasive, experimenter bias could not operate on participants' responses if the researcher could not distinguish between those who received the ad copy attributed to this source from those whose copy was attributed to other sources.

Mechanized procedures. Whereas this variant of the blind procedure can be employed in many different experimental settings, it cannot be used in all situations. In some settings, the experimental and control conditions cannot be studied simultaneously in the same location. In such cases, mechanization of the experimental procedures can provide the solution to the problem of experimenter expectancy. Instructions, manipulations, and procedures could be presented to participants via audiotape, videotape, or a computer laptop, with respondents' answers collected using these electronic devices, and data analyzed by an impartial computer. Procedures where data are "untouched by human hands" help render implausible the alternative hypothesis of experimenter expectancy, if these devices are employed intelligently and not subverted by inappropriate actions on the part of the investigator.

As experiments administered via computer technologies usually involve using a mouse (or touchscreen) and a keyboard as a proxy for social interactions, such procedures can prove uninvolved, uninteresting, or unreal to the participant. These reactions in turn can give rise to a

host of contaminating features. This need not be the case. With adequate preparation, a realistic, interesting, and sometimes even educational experimental interaction can be devised. And, as recent experience with computer-based experimental administration has demonstrated, the strategic use of computers to present experimental treatments, or to monitor participants' reactions, can be a positive and important step in the solution of the problem of experimenter bias.

Clearly, any list of suggestions for controlling or minimizing experimenter expectancy effects will prove incomplete. Many other approaches to the solution of the expectancy problem are possible but are idiosyncratic to the experimental setting employed. The important consideration to keep in mind is that attempts to prevent expectancy effects must be a part of a good experimental design. Double-blind procedures and mechanization should be used when possible.

Three Faces of External Validity

External validity refers to the question of whether an effect (and its underlying processes) that has been demonstrated in one research setting would be obtained in other settings, with different research participants and different research procedures. External validity is not a single construct but represents a whole set of questions about generalizability, each with somewhat different implications for the interpretation and extension of research findings (Brewer & Crano, *in press*). The sections that follow discuss three of the most important forms of external validity: robustness, ecological validity, and relevance. Each of these raises somewhat different questions about where, when, and to whom the results of a particular research study can be generalized.

Robustness: Can It Be Replicated?

A result that is replicable across a variety of settings, persons, and historical contexts is described as robust. In its most narrow sense, **robustness** is concerned with the external validity question of the extent an effect obtained in one laboratory can be replicated in another, at different times and locations and different participants. More broadly, the question has to do with whether the general effect holds up in the face of variations in participant populations and settings. Some findings appear fragile, obtainable only under highly controlled conditions in a specific context; other results hold up across significant variations in contexts under which they were originally established.

Technically, robustness would be demonstrated if a study were conducted using a randomly selected sample of participants from a broadly defined population, across a random sampling of settings. This approach to external validity implies that the researcher must have defined the populations and settings to which the effect of interest is to be generalized, and then must compile a listing of the populations and settings from which a sample is drawn. Such sampling designs, however, are usually impractical and can be expensive. More often, this form of generalizability is established by repeated replications in systematically varied settings and types of research participants. For instance, a finding initially demonstrated in a laboratory with college students from an eastern college in the United States may later be replicated with high school students in the Midwest and among members of a community organization in New England. Such replication strategies involving disparate samples are not only more practical, but they also have potential advantages for theory-testing purposes. If findings do not replicate in systematically selected samples, we sometimes gain clues as to what demographic factors may be moderators of the effect in question (Petty & Cacioppo, 1996).

Generalizability across multiple populations and settings should be distinguished from generalizability to a particular population. A phenomenon that is robust in the sense that it holds

up for the population at large may not be obtained for a specific subpopulation or in a specific context. If the question of generalizability is intended to extend to a specific target population (say for young adults, 19–25 years of age), then replications must be undertaken with samples drawn from that specific population of interest.

External validity, or generalizability, is related to settings as well as participant populations. The external validity of a finding is challenged if the relation found between independent and dependent variables is altered when essentially the same research design and procedures are conducted in a different laboratory or field setting or by experimenters with different characteristics. For example, Milgram's (1974) initial studies of obedience were conducted in a research laboratory at Yale University, but using participants recruited from the community of New Haven, Connecticut. Even though these experiments were conducted with a nonstudent sample, a legitimate question is the extent to which his findings would generalize to other settings. Because participants were drawn from outside the university and because many had no previous experience with college, the prestige and respect associated with a research laboratory at Yale may have made the participants more susceptible than they would have been in other settings to the demands for compliance Yale's reputation might have fostered.

To address this issue, Milgram replicated his experiment in a very different physical setting. Moving the research operation to a "seedy" office in the industrial town of Bridgeport, Connecticut, and adopting a fictitious identity of a psychological research firm, Milgram sought to minimize the reputational factors inherent in the Yale setting. In comparison with data obtained in the original study, the Bridgeport replication resulted in slightly lower but still dramatic rates of compliance to the experimenter. Thus, setting could be identified as a minor, but not crucial contributing factor to the original research findings.

Ecological Validity: Is It Representative?

The question of whether an effect holds up across a wide variety of people or settings is somewhat different from asking whether the effect is representative of what happens in everyday life. This is the essence of **ecological validity** – concerned with the external validity issue of the extent an effect occurs under conditions that are typical or representative in the population. The concept of ecological validity derives from Brunswik's (1956) advocacy of "representative design" in which research is conducted with probabilistic samplings of people and situations.

Representativeness is not the same as robustness. Robustness asks whether an effect can occur across different settings and people; ecological validity asks whether it does occur in the world as is. From the Brunswik perspective, findings obtained with atypical samples (e.g., college students) in atypical settings (e.g., the laboratory) never have ecological validity until they are demonstrated to occur naturally in more representative circumstances.

Many researchers (e.g., Berkowitz & Donnerstein, 1982; Mook, 1983; Petty & Cacioppo, 1996) take issue with the idea that the purpose of most research is to demonstrate that events actually occur in a particular population. They argue that testing a causal hypothesis requires demonstrating only that manipulating a cause can alter an outcome (i.e., an effect). Even most applied researchers are more interested in questions of whether interventions *can* change outcomes rather than what *does* happen under existing conditions. Ecological validity is too restrictive a conceptualization of generalizability for research that is designed to test causal hypotheses. Ecological validity is, however, crucial for research that is undertaken for descriptive or demonstration purposes.

Further, the setting in which a causal principle is demonstrated does not necessarily have to resemble the physical settings in which that principle operates in real life for the demonstration

to be valid. As Aronson et al. (1998) put it, most social psychology researchers are aiming for “psychological realism” rather than “mundane realism” in their experiments. An experimental setting may not resemble features in the real world, but still may capture processes that underlie events in the real world.

Relevance: Does It Matter?

In a sense, the question of ecological validity is also a question of **relevance**, which concerns the external validity issue of the extent an effect obtained is pertinent to events or phenomena that actually occur in the real world. However, relevance also has a broader meaning of whether findings are potentially useful or applicable to solving problems or improving quality of life. Relevance in this sense does not necessarily depend on the physical resemblance between the research setting in which an effect is demonstrated and the setting in which it is ultimately applied. Perceptual research on eye-hand coordination conducted in tightly controlled, artificial laboratory settings has proved valuable to the design of instrument panels in airplanes even though the research laboratory usually is not mistaken for a cockpit.

Relevance is the ultimate form of generalization and differences among research studies in attention to relevance is primarily a matter of degree rather than of kind. All social research is motivated ultimately by a desire to understand real and meaningful social behavior. But often the connections between basic research findings and their application are indirect and cumulative, rather than immediate. Relevance is a matter of social process – the process of how research results are transmitted and used rather than what the research results are (Brewer, 1997).

Is External Validity Important?

External validity – like other issues of validity – must be evaluated with respect to the purpose for which research is being conducted. When the research agenda is essentially descriptive, ecological validity may be essential. When the purpose is utilitarian, robustness of an effect is particularly critical. The fragility and nongeneralizability of a finding may be fatal if one’s goal is to design an intervention to solve some applied problem. On the other hand, it may not be so critical if the purpose of the research is testing explanatory cause-and-effect relationships, in which case internal validity is more important than satisfying the various forms of external validity.

In physics, for example, many phenomena can be demonstrated empirically only in a vacuum or with the aid of supercolliders. Nonetheless, the findings from these methods often are extremely important in developing an understanding of basic principles and ultimate application of the science. Mook (1983) has argued compellingly that the importance of external validity has been exaggerated in the psychological sciences. Most experimental research, he contends, is not intended to generalize directly from the artificial setting of the laboratory to “real life,” but to test predictions based on theory. He draws an important distinction between “generality of findings” and “generality of conclusions,” and holds that the latter purpose does not require the conditions of testing resemble those of real life. Understanding the underlying processes, not the nuances of the findings, is emphasized in Mook’s (1983) discussion of external validity.

In effect, Mook argued that construct validity is more important than forms of external validity when we are conducting theory-testing research. Nonetheless, the need for conceptual replication to establish construct validity requires robustness across research operations and settings. This requirement is similar to that for establishing external validity. The type of

systematic, programmatic research that accompanies the search for external validity inevitably contributes to the refinement and elaboration of theory as well as generalizability.

Conclusion

Issues surrounding the external validity of laboratory experiments are numerous and varied. We agree with the nearly unanimously accepted contention that the question of external validity, or generalizability, can never be answered definitively. However, this does not preclude experimenters from enhancing generalizability in their studies. For example, simply ensuring that a wide range of individuals are used in research means that replicated findings are not tied to a specific participant group (e.g., college sophomores). Understanding and offsetting, to the extent possible, the constraints that the experimental context places on participants also supports the external validity of research results. Conducting studies in ways that limit experimenter bias and experimental demand helps ensure the validity – external as well as internal – of experimental findings. And finally, dealing with issues of relevance to the participants, issues that matter, that are studied in realistic contexts, in experiments that produce strong effects all add confidence in the external validity of research results. These requirements are not easily met, but their satisfaction results in findings that are relevant and that more accurately reflect the relations that exist in the contexts in which they were investigated.

Notes

- 1 This would involve hiring experimenters or data analysts; the individual responsible for mounting the study could hardly be expected to do so without knowing his or her own hypothesis.
- 2 This observation illustrates the important conceptual distinction between methodological versus statistical control. A methodological control of expectancy would, through appropriate experimental design procedures, render the contamination of results by “experimenter expectancy” improbable. Conversely, statistical control would not rule out the possibility of the occurrence of experimenter effects, but rather would enable us to gauge their strength.

References

- Adair, J. G., & Epstein, J. S. (1968). Verbal cues in the mediation of experimenter bias. *Psychological Reports*, 22(3, PT. 2), 1045–1053.
- Ambady, N., Koo, J., Rosenthal, R., & Winograd, C. H. (2002). Physical therapists’ nonverbal communication predicts geriatric patients’ health outcomes. *Psychology and Aging*, 17(3), 443–452.
- Aronson, E., Wilson, T., & Brewer, M. B. (1998). Experimentation in social psychology. In D. Gilbert, S. Fiske, & G. Lindsey (Eds.), *The handbook of social psychology* (4th ed., Vol. 1, pp. 99–142). Boston, MA: McGraw-Hill.
- Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, 71, 230–244.
- Beecher, H. K. (1966). Pain: One mystery solved. *Science*, 151, 840–841.
- Berkowitz, L., & Donnerstein, E. (1982). External validity is more than skin deep: Some answers to criticisms of laboratory experiments. *American Psychologist*, 37, 245–257.
- Brewer, M. B. (1997). The social psychology of intergroup relations: Can research inform practice? *Journal of Social Issues*, 53, 197–211.
- Brewer, M. B., & Crano, W. D. (in press). Research design and issues of validity. In H. T. Reis, T. West, & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (3rd ed.). Cambridge: Cambridge University Press.
- Brunswick, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley: University of California Press.

- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally & Company.
- Carlopio, J., Adair, J. G., Lindsay, R. C. L., & Spinner, B. (1983). Avoiding artifact in the search for bias: The importance of assessing subjects' perceptions of the experiment. *Journal of Personality and Social Psychology*, 44, 693–701.
- Cemalcilar, Z., Baruh, L., Kezer, M., Soto, C. J., & John, O. P. (2021). Testing the BFI-2 in a non-WEIRD community sample. *Personality and Individual Differences*, 182, 111087.
- Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist*, 30, 116–127.
- de la Fuente-Fernández, R., Ruth, T. J., Sossi, V., Schulzer, M., Calne, D. B., & Stoessl, A. J. (2001). Expectation and dopamine release: Mechanism of the placebo effect in Parkinson's disease. *Science*, 293(5532), 1164–1166.
- Fode, K. L. (1960). *The effect of nonvisual and nonverbal interaction of experimenter bias*. Unpublished Master's Thesis, University of North Dakota.
- Gergen, K. J. (1973). Social psychology as history. *Journal of Personality and Social Psychology*, 26, 309–320.
- Gergen, K. J. (1976). Social psychology, science and history. *Personality and Social Psychology Bulletin*, 2, 373–383.
- Harris, M. J. (1991). Controversy and cumulation: Meta-analysis and research on interpersonal expectancy effects. *Personality and Social Psychology Bulletin*, 17(3), 316–322.
- Harris, M. J., & Rosenthal, R. (1985). Mediation of interpersonal expectancy effects: 31 meta-analyses. *Psychological Bulletin*, 97, 363–386.
- Harris, M. J., & Rosenthal, R. (1986). Four factors in the mediation of teacher expectancy effects. In R. S. Feldman (Ed.), *The social psychology of education: Current research and theory* (pp. 91–114). New York: Cambridge University Press.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33, 61–135.
- Jastro, J. (1900). *Fact and fable in psychology*. Boston, MA: Houghton Mifflin.
- Kjaervik, S. L., & Bushman, B. J. (2021). The link between narcissism and aggression: A meta-analytic review. *Psychological Bulletin*, 147, 477–503.
- Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, 67, 371–378.
- Milgram, S. (1965). Some conditions of obedience and disobedience to authority. *Human Relations*, 18, 57–75.
- Milgram, S. (1974). *Obedience to authority*. New York: Harper & Row.
- Moll, A. (1898). *Hypnotism* (4th ed.). New York: Scribner's.
- Mook, D. G. (1983). In defense of external invalidity. *American Psychologist*, 38, 379–387.
- Orne, M. (1962). On the social psychology of the psychological experiment. *American Psychologist*, 17, 776–783.
- Petty, R. E., & Cacioppo, J. T. (1996). Addressing disturbing and disturbed consumer behavior: Is it necessary to change the way we conduct behavioral science? *Journal of Marketing Research*, 33, 1–8.
- Pierce, A. H. (1908). The subconscious again. *Journal of Philosophy, Psychology, and Scientific Methods*, 5, 264–271.
- Reed, C. F., & Witt, P. N. (1965). Factors contributing to unexpected reactions in two human drug-placebo experiments. *Confinia Psychiatrica*, 8, 57–68.
- Rosenthal, R. (1966). *Experimenter effects in behavioral research*. New York: Appleton-Century-Crofts.
- Rosenthal, R. (2002a). Covert communication in classrooms, clinics, courtrooms, and cubicles. *American Psychologist*, 57(11), 839–849.
- Rosenthal, R. (2002b). Experimenter and clinician effects in scientific inquiry and clinical practice. *Prevention & Treatment*, 5, Article 38.
- Rosenthal, R. (2003). Covert communication in laboratories, classrooms, and the truly real world. *Current Directions in Psychological Science*, 12(5), 151–154.
- Rosenthal, R., & Fode, K. L. (1963). Three experiments in experimenter bias. *Psychological Reports*, 12, 491–511.
- Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom*. New York: Holt, Rinehart & Winston.
- Rosenthal, R., & Lawson, R. (1964). A longitudinal study of the effects of experimenter bias on the operant learning of laboratory rats. *Journal of Psychiatric Research*, 2, 61–72.
- Rosenthal, R., & Rosnow, R. W. (1975). *The volunteer subject*. New York: Wiley-Interscience.

- Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences*, 1, 377–415.
- Schlenker, B. R. (1974). Social psychology and science. *Journal of Personality and Social Psychology*, 29, 1–15.
- Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, 51, 515–530.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton: Mifflin and Company.
- Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science*, 24, 1875–1888.
- Sivacek, J. M., & Crano, W. D. (1982). Vested interest as a moderator of attitude–behavior consistency. *Journal of Personality and Social Psychology*, 43, 210–221.
- Strohmetz, D. B. (2008). Research artifacts and the social psychology of psychological experiments. *Social and Personality Psychology Compass*, 2, 861–877.
- Wallach, L., & Wallach, M. A. (1993). Gergen versus the mainstream: Are hypotheses in social psychology empirical? *Journal of Personality and Social Psychology*, 67, 233–242.
- Webber, S. J., & Cook, T. D. (1972). Subject effects in laboratory research: An examination of subject roles, demand characteristics, and valid inference. *Psychological Bulletin*, 77, 273–295.

8

CONDUCTING EXPERIMENTS OUTSIDE THE LABORATORY

In the preceding chapters, we discussed ways experiments conducted in laboratory environments could be made as realistic, involving, and impactful as possible. Even when such methods are used, however, it still is desirable to move outside the laboratory into field contexts to extend the construct and internal and external validities of the results of any research program. As stated throughout this book, research should be viewed as a process in which initial observation of a phenomenon gives rise to theory development, which is followed by empirical research. Discrepancies between hypotheses and research results lead us back to further observations, to theory revision and refinement, and then to additional research, as shown in Figure 8.1. Neither laboratory nor field research is accorded primacy in this diagram. Each has its place, and each should be engaged only after appropriate observation and theory development. In this chapter, we support this view by considering ways experimental methods can be applied outside the laboratory, and how laboratory and field experiments can be used in conjunction with each other. In later chapters, we will take up nonexperimental research methods in field contexts.

The distinction between laboratory and field is one of setting, the context in which the research is conducted. A laboratory is a designated location where participants must go to engage in the research. A **field study** is research conducted outside the laboratory, in which the researcher administers the study in participants' own naturalistic environments. With the advent of the Internet as a venue for research, we move the experiment even farther away from the sterile confines of the laboratory to a virtual location, completed in the convenience of participants' own worlds, often in their homes or offices.

The distinction between lab and field is not always clear-cut. The school classroom, for instance, is sometimes converted into a "laboratory" for research purposes, and many laboratory studies have been conducted under the guise of some other activity, such as job interviews or group discussions. However, in general the feature that separates laboratory and field experiments is participants' awareness that they are involved in a scientific study. In field settings, even when informed in advance that a study is underway, participants are less likely to be as conscious of their behaviors and the research goals as they are in a laboratory environment. Studies conducted on the Internet can be more like laboratory experiments if respondents log on to a site knowing they are participating in a scientific investigation, or more like a field study if the research is being conducted using naturally occurring forums such as chat rooms or other

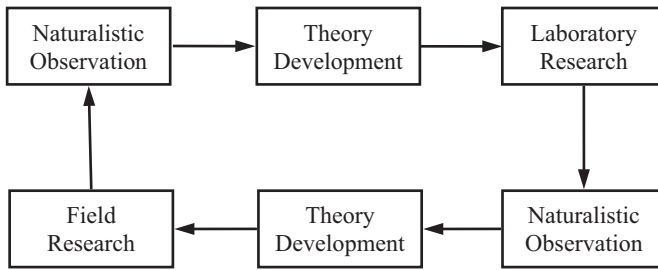


FIGURE 8.1 The research cycle, in which naturalistic observation leads to theory development, which leads to laboratory or field research, leading to more naturalistic observation, theory development, laboratory or field research in a continuous cycle of knowledge accretion.

social media sites as the source of data (see Smith & Boster, 2009; Sun & Lu, 2022; Walther et al., 2010).

As a general research strategy, the movement between laboratory and field settings should involve a two-way street. Many phenomena of interest to social scientists are first observed in naturalistic settings, but in the complex environment of the real world, it often is difficult to pinpoint the specific effect one variable has on another. To test a particular cause-and-effect relationship, the independent variable(s) of interest must be disembedded from natural contexts or causal networks, and translated into appropriate operations to be studied in the laboratory. The laboratory is an ideal setting for the experimenter to take special precautions to control and prevent the intrusiveness of extraneous variables from becoming confounds. Once a particular causal relationship has been established in the experimental laboratory or Internet experiment, the reasons for moving back into the field as a research site are exactly the reverse of those we have given. It is important to re-embed the causal variable into its natural context, to be certain that its effect is not suppressed or reversed when the phenomenon normally occurs.

Research Settings and Issues of Validity

The interchange between laboratory and field experiments is critical if we are to establish the external and construct validity of a research finding. For the most part, internal validity is inherent in a single study. With sufficient information about how an experiment was conducted, how participants were assigned to treatments, and how treatments were enacted, we should be able to assess whether the results of a particular study are internally valid. However, issues involving construct validity or external validity can rarely be resolved within the context of a single experiment. These questions require multiple replications of the effect, ideally across different research contexts, before meaningful assessments are possible.

It is important to distinguish between two types of replication research, where replication refers to the design and conduct of a new study designed to investigate the replicability of results of an earlier one. In an **exact replication**, an attempt is made to reproduce the results of a previous study by using the same procedures, particularly the same operationalizations, to represent the same constructs. That is, the operational definitions and conceptual definitions are identical to those of the original study. Only the participants, the time, and the place (and, usually, the researcher) are changed. The purpose is to determine whether a result can be repeated under slightly different circumstances. In a **conceptual replication**, an attempt is made to reproduce

the results of a previous study by using different operational definitions to represent the original constructs. To establish external validity of a research result, it is sufficient to demonstrate that the same independent variable has a similar effect on the dependent variable in different contexts with different types of participants. To establish construct validity, conceptual replications are required, in which the operationalizations of variables are *different* from those of the original study.¹

In principle, exact replications change the contextual environment of the studies while holding research procedures constant. Conceptual replications change both context and procedures. Variations in context and research operations are neither independent nor mutually exclusive, and both are closely related to alterations in the setting in which an experiment is conducted. In many cases, it is difficult to change the context in which a study takes place without altering the entire research setup, and this is particularly true when a replication involves the shift from laboratory to field or Internet settings. The advantages of field experiments are best realized when the operations of the independent and dependent variables are translated to be appropriate to the new context. Such modifications often involve a fundamental rethinking of the theoretical variables and a concern with conceptual, rather than exact replication of the original study.

Although specific procedures may differ, the rationale for the conduct of either field or laboratory experiments is the same. For purposes of the present chapter, however, we will pay particular attention to the ways in which the conduct of a field or Internet experiment is most likely to differ from that of the typical laboratory experiment.

Constructing a Field Experiment

Selection of Participants

Field experiments generally differ from laboratory experiments in the extent to which participants are consciously aware that they are involved in a research project. The two settings (field and laboratory) also differ in the types of people who usually are recruited as participants. A major criticism of laboratory research has been the extent to which it relies on college students as participants (see Krantz & Dalal, 2000; Sears, 1986). Field experiments often provide for a broader, or at least different representation of participant populations, and these participant variations are even more evident in Internet research. Researchers sometimes select particular field sites specifically to reach participant groups of a particular age or occupation – they go to aged-care homes to study the elderly, to schools to study young children, to circuses to study clowns, and to courtrooms to study lawyers. Other settings (e.g., the city streets) do not permit such selection over a narrowly defined group, but do provide for a wider demographic range of potential participants. It should be emphasized, however, that moving into field settings does not automatically guarantee greater representativeness of participants. Anderson et al. (1999) conducted a comparative review of laboratory and field research in various areas of social psychology. In general, correspondence was high between results from lab- and field-based studies involving conceptually similar independent and dependent variables. The authors concluded that “the psychological laboratory has generally produced psychological truths, rather than trivialities” (p. 3).

Control Over the Independent Variable

Although the essence of experimentation is systematic manipulation by the researcher of variations in treatments or conditions that constitute the independent variable, the extent of

experimenter-controlled manipulation in different research settings is a matter of degree. In some instances, the researcher creates experimental conditions from scratch, controlling background context as well as experimental variations (Levine et al., 2002). In other cases, the experimenter controls less of the setting but introduces some systematic variation into the existing conditions, as in the field experiment by Piliavin et al. (1969), where the behavior of an experimental confederate in a subway train was manipulated to study bystander helping in that setting.

In other field research contexts, the experimenter neither manipulates the stimulus conditions directly nor controls participant attention, but instead selects from among naturally occurring stimuli in the field that represent the independent variable of interest. Here the line between experimental and nonexperimental research becomes thin indeed, and the distinction depends largely on how well the selected field conditions can be standardized across participants. A good illustration of the use of laboratory research in conjunction with selected field sites comes from the literature on mood and altruism. Many different mood induction manipulations have been developed in laboratory settings. Typically, they require participants to read emotionally positive or negative passages or to reminisce about happy or sad experiences in their own past (Forgas, 2000). Following mood induction, participants are given an opportunity to respond to some form of request – to help a person in need, to exhibit generosity, and so on. Results generally show positive moods elevate prosocial behavior compared to negative moods. Despite multiple replications of this effect in different laboratories with different investigators, however, these findings have been challenged both because of the artificiality of the settings in which the behaviors are induced, and because of the potential demand characteristics associated with the rather unusual laboratory procedures used to induce moods.

To counter these criticisms, researchers took advantage of a natural mood induction situation based on the emotional impact of selected motion pictures (e.g., Underwood et al., 1977). Following pilot research in which ratings were obtained from moviegoers, a double feature consisting of “Lady Sings the Blues” and “The Sterile Cuckoo” was selected for its negative mood-inducing qualities for use as the experimental condition. Two other double features were selected as the control condition to induce a neutral mood. A commonly occurring event – solicitation of donations to a nationally known charity with collection boxes set up outside the movie theater lobby – was used to operationally define the dependent variable of helping behavior.

Having taken the laboratory mood induction manipulation and translated it into a field manipulation, a major design problem the researchers then encountered was that of participant self-selection to the type of movie preferred. Whereas random assignment of volunteer moviegoers to the two types of movies was a possibility, such a strategy would have recreated many of the elements of artificiality and reactivity that the field setting was designed to avoid. Therefore, the investigators decided to live with self-selection and to alter the research design to take its effect into consideration. For this purpose, timing the collection of charitable donations at the various movie theaters was randomly alternated across different nights, to occur either while most people were entering the theater (prior to seeing the movies) or while leaving it (after seeing the double feature). The rate of donation of arriving moviegoers could be used as a pretest check of preexisting differences between the two groups prior to the mood induction. Fortunately, there were no differences in pre-movie donation rates as a function of the type of movie the participants chose to see, whereas post-movie donations differed significantly in the direction of lowered donation rates after watching the sad, but not neutral movies.

Two points should be emphasized with respect to this illustration of field research. First, the field version of the basic research paradigm was not and could not be simply a “transplanted” replication of the laboratory operations. Significant alterations were necessary to take full

advantage of the naturalistic setting. The researchers had considerably less control in the field setting. They could not control the implementation of the stimulus conditions or extraneous sources of variation. On any given night, a host of irrelevant events may have occurred during the course of the movies (e.g., a break-down of projectors, a disturbance in the audience, a mass shooting in a neighboring city) that could have interfered with the mood manipulation. The researchers were not only helpless to prevent such events, but might not even have been aware of them if they did occur. In addition, the experimenters were unable to assign participants randomly to conditions in this field setting, and had to rely on luck to establish equivalence in initial donation rates between the two groups.

The second point to be emphasized is that the results of the field experiment as a *single isolated study* would have been difficult to interpret without the results obtained from conceptually related laboratory experiments. This difficulty is partly due to the ambiguities introduced by the alterations in design, and partly to the constraints on measurement inherent in the field situation where random assignment usually is not feasible. The concurrence of results in laboratory and field settings greatly enhances our confidence in the findings obtained from both sets of operationalizations of mood in measuring the same underlying concept. Had the field experiment failed to replicate the laboratory results, however, many possible alternative explanations for this discrepancy would have arisen, and would have rendered interpretation very difficult.

Random Assignment in Field Settings

Participant self-selection problems plague field experimentation in many ways. In the field experiment on the relation between mood and helping behavior, discussed earlier, random assignment to conditions was not even attempted. Instead, the effects of potential selection factors were handled in another way, which involved an element of risk taking. The pre-movie data collection served as a pretest check on the assumption that people who attend sad movies were not inherently different from people attending other movies in their propensity to give to charities. But what if that assumption had proved false and there had been an initial difference in the rate of donations from attendants at the different types of movies? Such prior baseline differences in behavior would have made interpretation of donation differences after movie exposure haphazard at best. In this case, the researchers were taking a gamble in counting on the absence of pre-movie differences in personal characteristics between attendees in movies. This is not a safe bet. Personal experience, or better yet, pilot research, could have led the researchers to expect that factors determining the type of movie that attracted were relevant or irrelevant to their propensity to give to charity.

In other research contexts, too, research may rely inadvertently on the essentially haphazard distribution of naturally occurring events as equivalent to controlled experimental design. Parker et al. (1980), for instance, studied the outcomes of a natural disaster – a devastating brush fire in a Southern California community – on the premise that the pattern of destruction of private homes in the fire constituted a “natural randomization” process. Among homes in close proximity at the height of the fire, only chance factors – shifts in wind direction and velocity, location of firefighting equipment, and traffic congestion – determined which structures were burned to the ground and which remained standing when the fire was brought under control. Thus, homeowners who were victims of the fire and those who were not could be regarded as essentially equivalent prior to the effects of the fire, and differences in their attitudes and perceptions following the fire could be attributed to their post-fire experience. As this “assignment” was made by Mother Nature rather than the researchers, the study is most properly considered a quasi-experiment (see Chapter 9). When comparisons are made between such naturally selected

groups, the burden of proof rests on the investigator to make a convincing case that the groups were not likely initially to differ systematically in any relevant dimensions other than the naturally arising event of interest.

One should not conclude from these examples that experimenter-controlled random assignment is not possible in field experiments. Often, the nature of the experimental manipulation is such that the researcher can deliver different versions or conditions to potential participants in accord with a random schedule. Consider a classic study of helping behavior in which the effect of positive, rather than negative, mood was being investigated. As one manipulation of positive mood, Isen and Levin (1972) arranged it so that some users of a public phone booth would find a coin in the return slot of the telephone as they started to make their call. Although the researchers had no control over which persons would make use of the targeted phone booths during the course of the experiment, they could control the timing and frequency that coins were or were not placed in the return slot. They alternated these conditions on a random basis and then observed the behavior of the next caller who happened to use the selected phone booth. With this kind of random assignment, in which conditions are randomly assigned to participants rather than participants randomly assigned to conditions, the researchers could be relatively confident that no prior-existing participant differences influenced their results.

In some field research efforts, the investigator may be able to assign participants randomly to conditions, but once assigned some participants may fail to participate or to experience the experimental manipulation. If such self-determined participant attrition occurs differentially across treatment conditions, the experimental design is seriously compromised. One way of preserving the advantages of random assignment in such cases is to include participants in their assigned experimental conditions for purposes of analysis regardless of whether they were exposed to the treatment or not (assuming, of course, we are able to obtain measures on the dependent variable for these participants). This is called an *intention-to-treat* approach, and is used widely in medical research (Katz et al., 2020; Kochovska et al., 2020).

Freedman and Fraser (1966) applied this method to two field experiments to test the effectiveness of the “foot-in-the-door” technique for enhancing compliance to a rather large, intrusive request from the researcher (i.e., to permit a five-person market survey team to come into one’s home for two hours to classify household products). Of primary interest was the rate of compliance to this large request by participants who had been contacted previously with a small request (e.g., to respond to a brief market survey over the telephone), in comparison to control participants who were not contacted until the time of the larger request.

The purpose of the manipulation in Freedman and Fraser’s study was to test the effect of compliance to the initial small request on responses to the later one. However, nearly one-third of the participants in the experimental group (those who were assigned to receive the initial small request) did not comply. This posed a logical problem for the experimenters. How were they to treat these participants’ data? Although the participants had been assigned randomly to the experimental condition, they did not actually perform the action that was presumed to be the cause of acquiescence to a later, larger request. However, if they had removed the participants from the analysis, the random assignment process would have been compromised, and the comparability of the two groups would be in serious doubt.

Freedman and Fraser’s solution was proper: they included responses from all participants in the originally assigned groups, regardless of their compliance with the initial request. This was a conservative decision, because the full treatment was significantly diluted in the experimental group. As it turned out, the initial compliance effect was powerful enough to generate a significant difference between the two groups (on the order of 50% vs. 20% compliance), *despite* the dilution of the experimental condition. If the results had been more equivocal, however, we

would have been uncertain whether to attribute the absence of significant differences to lack of treatment effects or to failure to achieve the experimental manipulation.

Assessing Dependent Variables in Field Settings

In many field contexts, the design and evaluation processes are identical to that of laboratory experiments. In the guise of a person-on-the-street interview or a market research survey, for example, field researchers may elicit self-reports of the attitudes, perceptions, judgments, or preferences of randomly selected individuals. Or, measures may be designed to assess people's willingness to engage in relevant acts such as signing a petition or committing themselves to some future social or political cause. Finally, situations may be constructed to elicit the type of behavior of interest to the experimenter, such as providing participants with opportunities to donate to charity. One advantage of experimentation in field settings is the potential for assessing *behaviors* that are, in and of themselves, of some significance to the participant. Instead of asking participants to *report* on perceptions or intentions, we observe them engaging in behaviors with real consequences. In such cases, our dependent measures may be less likely influenced by experimental demand characteristics or social desirability response biases. In laboratory settings, participants might check a particular point on an evaluation scale to please the experimenter or to look good; but we think very few people would choose to engage in a difficult or onerous task in a field study unless there were more meaningful reasons to do so.

Unobtrusive Measures

In some field settings, the kinds of dependent measures typically employed in laboratory studies might be viewed as so intrusive that they would destroy the natural flow of events. For this reason, field experiments often are characterized as using concealed or indirect measures of the dependent variable under study.

In a sense, most measures of psychological variables are indirect, in that we have no direct access to the thoughts or perceptions of another person. However, some measures are more indirect than others. Indirect measures are those for which the link to the concept of interest involves a hypothetical intervening process. For example, in an interesting study of the "illusion of control" over chance events, Langer (1975) sold 50-cent lottery tickets to participants under one of two conditions. In one condition, participants were handed a particular ticket by the salesperson (experimenter); in the other, participants were allowed to choose their own tickets from the available set. Regardless of condition, each ticket had an equal probability of winning the lottery. Langer was interested in examining the effect of the illusory "control" implied in the choice condition on participants' confidence that they might possess the winning ticket.

Rather than simply asking the participants how confident they felt, Langer used a less direct measure of this variable. Each participant was approached after purchasing a ticket, but before the lottery, and was told that someone else wanted to purchase a ticket and the seller had run out. Participants then were asked how much they would be willing to sell back their tickets. The reasoning behind this procedure was that the asking price requested by the participant would reflect the ticket's subjective value, which reflected the perceived probability the participant attached to their ticket winning the lottery. As predicted, those who were allowed to choose their own tickets wanted significantly more money to sell their ticket back than participants who had been given their tickets with no choice.

What is interesting about this use of an indirect behavioral measure is the likelihood that participants might have been too embarrassed to report on their differential confidence had they

been asked directly whether they thought they had a winning ticket; after all, they would know that the “objective” view would be that the probability was quite low and attributable purely to chance. Assuming the indirect measure used was closely related to true subjective confidence, it may have detected an effect that would not have appeared in the results of direct self-report.

Indirect measures are among a variety of techniques used by field researchers to make *unobtrusive measurements* of the dependent variable of interest (see Webb et al., 1981). Some unobtrusive measures are based on observations of ongoing behavior, using methods of observation that interfere minimally or not at all with the occurrence of the behavior. For instance, voluntary seating aggregation patterns have been used as an index of racial behaviors under varied conditions of classroom desegregation; observational studies of conformity have recorded public behaviors such as pedestrians crossing against traffic lights; and studies of natural language often resort to eavesdropping on conversations in public places.

Cialdini and associates (1976) used naturalistic observation of clothing and accessories to study what they called the “Basking in Reflected Glory” phenomenon. They recorded the number of students wearing t-shirts and other apparel bearing the school name or insignia in introductory psychology classes at seven universities each Monday during football season. The proportion of students wearing such apparel at each school proved to be significantly greater on Mondays following a victory by that school’s team than on days following defeat. A simple monitoring of public displays provided quantitative confirmation of the hypothesized tendency to identify with success – or to bask in reflected glory. Sigelman (1986) noticed that those who supported the winning candidate tended to keep their political signs and posters up longer than those whose candidates lost. Further research has shown that people tend to assume electoral victors’ policies are more in line with their views than those of losers, even though these beliefs might not have been held *before* the election (Alabastro et al., 2013; Quist & Crano, 2003).

Other observational techniques may rely on the use of hidden hardware for audio or video recording of events that are coded and analyzed. In some circumstances, the ethics of such research have come under heavy criticism. Less questionable is research using “wearable technology” (e.g., smartphone, Fitbit, iWatch, etc.) to collect data in a semi-unobtrusive manner. The wearer is aware of the measurement device, but it soon becomes so much a part of normal use that it would not seem to interfere with natural behavior. A further advantage of such devices is their relatively low expense and their capacity to monitor mental and physical states continuously, often providing information of which the wearer is not aware. For example, Sequeira et al. (2020, p. 314) reviewed 829 studies of wearable tools used to investigate child and adolescent depression. Their research on the 30 studies that met their criteria found wearables captured a variety of data including depressive symptom monitoring.

Some unobtrusive techniques make use of the natural recording of events outside the experimenter’s control, such as physical traces left after an event. Ball’s (2020) research on desktop graffiti provides an interesting example of this approach. In 2007, Virginia Tech University was the scene of a catastrophic mass shooting. An undergraduate student at the school, Seung-Hui Cho, killed 32 people and wounded 17 more before committing suicide. Ball collected a sample of desks prior to the massacre, and photographed the graffiti written on each of them. The desks were then cleaned and returned to the classrooms. She resampled the same desks one year later, and after the mass shooting. The content-analyzed graffiti collected before and after the shooting served as repeated dependent measures. Content of the pre- and post-massacre did not differ much. The analyses disclosed little racial/ethnic prejudice, which might have been expected given the ethnicity of the shooter. The usual themes, love and sex, politics, religion, etc., were consistent across time.

Potential problems with unobtrusive measures must be considered when using them as dependent measures. A central issue has to do with measurement reliability. In general, the reliability of unobtrusive measures will not be as great as the more direct measures they are designed to mimic (Webb et al., 1981). As might be expected, therefore, the measurement validity of the dependent variable – the extent to which measures reflect what they are supposed to measure – also is likely to be of greater concern with unobtrusive measurements. This concern comes about because the farther removed the actual measure is from the concept of interest, the less likely it is to prove valid. Researchers have tried to offset reliability and validity issues by combining unobtrusive measurement studies with other forms of investigation (Medin et al., 2015), but the issue does not admit to an easy formulaic solution.

Field Experimentation and Application

Conceptual replication highlights the advantages of combining laboratory and field experimentation for purposes of theory-building, but the interplay between laboratory and field research is critical as well to the development of effective applications in social science. Basic experimental research may isolate important causal processes, but convincing demonstrations that those processes operate in applied settings are essential before theory can be converted into practice. The research literature on people's perceptions of control provides a particularly good example of how a synthesis between field and laboratory experiments can work at its best. This research began with animal research in the laboratory (Brady, 1958), extended to field studies of stress in humans (e.g., Janis, 1958; Langer & Rodin, 1976), then moved to laboratory analogues (e.g., Glass & Singer, 1972), and back to the field (e.g., Aspinwall & Taylor, 1992; Johnson & Leventhal, 1974; Mills & Krantz, 1979). Results from both types of settings repeatedly have demonstrated the potent effect of the perception of control or responsibility on an organism's ability to cope with stressful events (Moghaddam & Studer, 1998; Taylor & Aspinwall, 1996). Even the illusion that one has personal control over the onset or the consequences of potential stressors was apparently sufficient to increase tolerance for stress and reduce adverse effects. As a result of these findings, procedures applicable in medical practice and the administration of health care institutions have been developed for inducing actual or perceived personal control (e.g., Thompson & Collins, 1995). At the same time, the fact that field applications permit testing research hypotheses in the presence of severe, noxious, or potentially life-threatening situations has contributed substantially to our basic theoretical understanding of the role of psychological factors in physiological processes (Cortez & Bugenthal, 1995; Vancleef & Peters, 2011).

The Internet as a Site for Experimental Research

The advent of the Internet, and social networking sites in particular (e.g., TikTok, Twitter, Instagram, YouTube, etc.), has created a whole new venue for field research in the social sciences. Although the first studies conducted via the Internet are of recent vintage (some consider the survey study of Kiesler & Sproul, 1986, the first Internet study, and research by Krantz et al., 1997, among the first published Web-based experiments), the potential of the Internet as a research venue is enormous. The entire world is increasingly becoming wired, thus the Internet offers an opportunity to tap into it as a source of research participants as well as a site for observing social activity.² The use of the Internet as both a medium for conducting research studies and as a subject of research has grown accordingly. A review by Wilson et al. (2012) identified 412 research studies on Facebook, published between 2005 and 2012, demonstrating

that “researchers have discovered the utility of Facebook as a novel tool to observe behavior in a naturalistic setting, test hypotheses, and recruit participants” (p. 3). The use of social media to study social issues has increased exponentially as time has passed (e.g., Acerbi, 2022; Busam & Solomon-Moore, 2021; Canary et al., 2022; Delle et al., 2022).

At this point we need to make a distinction between research that uses the Internet to recruit respondents for participation in online surveys or experiments (also known as *crowdsourcing*) and research designed to study online behavior or its effects. Research employing online surveys will be covered in Chapter 13. Of more relevance to the present chapter is (a) research that uses the Internet as a site for conducting experiments and (b) research on how, when, and why people use online social networks as a form of social behavior.

Research on Social Network Activity

The review by Wilson et al. (2012) of research on Facebook illustrates how online activity can itself be studied as a form of social behavior in a naturalistic setting. Research has been conducted to address questions about who uses online social networks, why people make use of this form of social communication, the content of information that is disclosed online, and the consequences of online social activity for real-life relationships. Much of this research utilizes surveys (conducted offline) that assess respondents’ self-reported use of online social networking (and correlated variables), but some research uses online activity as data, including rates of usage of online sites over time, content of user profiles, and demographic statistics on users.

For the most part, the research conducted thus far has been descriptive or correlational, similar to observational research methods (see Chapter 13) or content analysis research (see Chapter 14). However, it is possible to introduce experimental interventions into ongoing network activity, such as randomly assigning participants to increase their frequency of status postings on Facebook to assess effects on perceived social connectedness (Deters & Mehl, 2013), or posting experimentally designed messages to online discussion groups to assess effects on emotional content of responses (Gonsalkorale & von Hippel, 2011; Walther et al., 2010). Such studies, then, are a special class of field experiments, with practical, methodological, and ethical considerations similar to experimental research in other field sites.

Conducting Experiments Online

Experiments on online behavior are still relatively rare, but the use of the Internet as a venue for conducting online data is becoming more and more common as software and specialized sites for such research become available. In many cases, versions of laboratory experiments are created for mounting online to determine whether findings from previous lab studies can be replicated in this venue, with new participant populations (e.g., Lutz, 2016; Simmons et al., 2022; Washington et al., 2019). Research by Williams et al. (2000) supplies an interesting example of this form of conceptual replication. Williams (1997) was interested in studying the effects of ostracism – being ignored and excluded from ongoing social interactions. Considerable research on ostracism has indicated that experiencing social ostracism is highly aversive (e.g., Rudert et al., 2019; Williams & Zadro, 2005).

Williams (1997; also see Hales & Williams, 2021) theorized that ostracism threatens four fundamental needs: belongingness, self-esteem, control, and meaningful existence. From this central axiom, Williams developed a model that predicts the psychological reactions that ensue in response to ostracism, a model tested in laboratory experiments in which confederates are trained to engage in a ball-tossing exchange while excluding the experimental participant.

To replicate these experiments online Williams et al. (2000) mounted a Web experiment in which participants engaged in an online game (called “cyberball”) with two other participants (actually, these others’ responses were programmed). Respondents were told that they should envision “throwing and catching a flying disc according to the images that were shown on their screen” (p. 751). The images systematically included the participants in the game, or systematically excluded them (where participants received the disk 33%, 20%, or never, after the first throw). Measurements taken immediately afterward showed that a greater degree of being excluded in this activity had greater aversive impact on participants, as predicted. Perceptions of belonging and self-esteem were most severely threatened by ostracism, even in this online version.

This research supported Williams’ (1997) model, and provides an interesting and useful model of a Web-based experiment. The study recruited 1,400 participants from 62 different countries, and they ranged from 13 to 55 years of age. The diversity and magnitude of the participant sample suggest that results on feeling socially rejected when being excluded are probably generalizable beyond the boundaries of the standard college laboratory or student sample. Perhaps even more important, having open-access sites like this affords researchers the possibility of obtaining extremely large and diverse samples, sometimes greater than 20,000 respondents, for study (see Johnson et al., 2014). Further research comparing data from Internet-delivered experiments conducted on Psych Experiments (a public online psychological laboratory) with comparable data from standard laboratory research indicates that Internet results mirror lab results across a wide range of psychological effects (McGraw et al., 2000).

Methodological Concerns: Reliability and Validity

Before considering some of the practical issues involved in using the Internet as a research venue, it makes sense to determine if it *should* be used. The Internet has a number of strong proponents (e.g., see Birnbaum, 2000a; Buhrmester et al., 2011; Reips, 2000), but even these researchers advise caution when mounting or interpreting Internet-based research. These warnings have been echoed across a number of domains, but the general consensus is that crowdsourced participants are no more problematic than those studied in the typical face-to-face laboratory, or respondents drawn from larger surveys (e.g., Casler et al., 2013; Hauser & Schwarz, 2016; McCredie & Morey, 2019; Necka et al., 2016).

The issues that should be considered when using online venues as subject sources are the same as those we face with the more traditional methods including measurement reliability and validity. When using the Web as a research modality, issues of reliability are the same as those we confront when using conventional laboratory methods. The measures we use in Internet research must be psychometrically sound, which is not different from the requirement of the in-person methods. In some ways, the Internet offers greater opportunities to develop measures and scales to assess measurement reliability and validity. Buchanan (2000) suggests that with its abundance of available research participants, and their diversity relative to the standard college sample, developing and refining measurement instruments might be facilitated by using the Internet. Many respondents often are needed in developing scales of fact or opinion, and the necessary sample size is readily available online, and sometimes unattainable in traditional research settings.

Assessing measurement validity, specifically the construct validity of measures using the Internet instead of traditional methods, is sometimes difficult. One common approach to understand possible differences between the two modalities is to conduct the investigation in the traditional laboratory and also on the Web. Investigators cite similarities in research findings

and scale reliability between the two methods to indicate the convergent validity of research results. Although this approach is appealing, it rests on a confirmation of the null hypothesis to establish (convergent) validity of different methods. Parallel results from laboratory and the Internet are encouraging, but may be attributable to a host of factors. Sometimes, instead of determining convergent validity of methods, the goal may instead be to determine the topics of investigation or the types of scales that may produce disparate findings between these two modalities of administration. For example, experimental research has found that when either a questionnaire was given to White participants inside a laboratory, they managed their impression by offering socially desirable responses that indicated greater favorability of Blacks over Whites (Evans et al., 2003). However, when researchers replicated the study on the Internet, out-group-favoring responses were no longer evident. In this case, administering the measure online was judged as producing less defensiveness, thereby prompting more valid responses.

Conducting Online Studies

A detailed roadmap on how to make full use of the Internet to facilitate research is beyond the scope of this book. However, a number of such books are available as are Internet sites to help facilitate social research (e.g., Birnbaum, 2000b; De Simone et al., 2015; Jones, 1999). Many of the particulars involved in conducting a research project, such as Institutional Review Board approval and the use of incentives, should not be neglected just because it is implemented online.

There are three general steps useful in designing practically any study to be conducted on the Internet. They involve finding programs that allow different treatments to be administered to research participants. The second step is to find programs to host and administer the study.³

As with any experimental study, the online research process involves designing the treatment materials and stimuli. Experimental treatments commonly employed in Internet research include manipulations of instructions, or presentation of vignettes, messages, photographs, and videos. Some types of experimental manipulations considered appropriate in laboratory settings lack realism or are impossible to implement when conducted on the Internet. Problematic or impossible conversion of formerly “live” treatments to an online format include environmental manipulations, such as having smoke piped into a room to determine participants’ reactions to an emergency, as an online equivalent would fail to capture the smell of the smoke or the urgency of the situation. On the other hand, measures that require a measure of latency are perhaps easier to use than in-person research contexts. The use of more complex manipulations, such as watching a video of a persuasive health message embedded in a website, may be problematic for participants due to technical constraints (Skitka & Sargis, 2006), but technological advances generally can resolve this issue (e.g., Donaldson et al., 2016).

Be aware that not everyone with Internet access has sufficient bandwidth to view multimedia presentations: Participants might be required to wait patiently for the stimuli of a complex treatment to download, and these awkward moments of staring at an inactive screen contribute to lower attention and higher participant dropout rates. Long download times also may lead participants to skip over the material and proceed immediately to the dependent variable questions. Although experimenters can design the webpage so that participants will not be able to proceed until fully downloaded, and can unobtrusively measure the time spent on a specific page on the screen, the possibility remains that participants may not have the necessary patience or bandwidth to view complex stimuli.

After designing the materials and stimuli for the treatment conditions, the second step involves finding a website capable of hosting and administering your online study. Given technical advances in freely available online-based programs, researchers no longer need to learn

website programming code (HTML) to conduct most types of research designs. Featuring intuitive point-and-click interfaces, the online programs noted earlier, and many others, allow researchers to upload treatment materials, randomize participants to study conditions, specify the presentation order of stimuli and measures, and vary the number of available response options, to name a few possibilities. A major advantage of using these online programs is that participant responses are automatically recorded into a database or spreadsheet, avoiding data entry errors, and easily downloaded and saved at the researcher's convenience.

Once the Web experiment is designed and placed on a website, the final step involves participant recruitment. Crowdsourcing websites allowing the public to sign up for various tasks for compensation are a potential venue for recruitment of large and diverse samples. As noted, Amazon's Mechanical Turk, Turk Prime, and their current "offspring," CloudResearch, are such sites popular with social researchers. Once an account is created and the website link to the study is indicated, the researcher offers a set payment amount for each potential participant. Participants make the decision of whether to participate based on the research description provided by the investigator and probably the incentives offered. Evaluation of data using this recruitment source has concluded that measures administered through Mechanical Turk yield responses highly similar to measures involving traditional methods of data collection. Amso et al. (2022) published a free volume of nearly 500 pages of studies demonstrating the utility of online research techniques (also see Buhrmester et al., 2011).

Using this and other crowdsourcing sites, it also is possible to conduct research longitudinally, as previous participants are automatically tracked by a non-identifiable code that might be used to invite these individuals to complete a subsequent round. Researchers also could opt to restrict the study to people from a particular state or country, or make it available internationally. Alternatively, researchers targeting a specific population may recruit participants through existing research panels or other online techniques. Potential members of a targeted group (e.g., people with attention deficit hyperactivity disorder) could be recruited by posting a link to the study on a message board frequented by those with this issue. In addition, e-mails can be sent to various mailing lists that might contain subscribers interested in the topic under investigation. Both recruitment techniques sometimes require establishing the rapport and trust necessary to gain entry, so that one may then be granted the privilege to post a message to recruit members.

Misconceptions about Internet Research

To determine the veracity of several popular notions about Internet research methods, a large-scale study involving 361,703 Internet participants was compared to the aggregation of participants from 510 published studies that used traditional data collection methods (Gosling et al., 2004). Based on this study, we can dispel four of the most widespread concerns about the use of the Internet for experimental research. First, although Internet samples are not entirely representative of the population, they are markedly more demographically diverse than investigations using traditional methods, like those employing undergraduate samples. Online participants tend to come from more diverse backgrounds and cultures – after all, the Internet can attract participants from all over the world (Germine et al., 2012).

Early research supported the stereotype that online participants were predominantly educated White males. Today, the Internet is now available in many public venues – libraries, schools, and coffee shops – and across different electronic devices such as cell phones, making it more accessible to the general populace than ever before. Although the gender gap in Internet use has largely disappeared, a small racial disparity remains, with White respondents being more likely to have access to this medium than Latino or Black respondents, usually attributed

to income differences. Still, the Internet remains impractical for recruiting from certain populations (e.g., very young, very poor, homeless, or illiterate people, etc.)

A second misconception of Internet users is that they are less well-adjusted than those who do not use the Internet. Gosling and colleagues (2004) found this was not true, and considerable research has supported this observation. Gosling and colleagues' Internet respondents were not drastically disparate from the usual college sample on indicators of social adjustment or mental health. Kraut and colleagues (1998) found the view of the typical Internet user as psychologically maladjusted, neglecting interpersonal relationships and social responsibilities for the sake of the sterile computer was largely prompted by early Internet research with convenience samples (Stoll, 1995; Turkle, 1996). The description is not applicable to the majority of today's online users. Like many, Bargh and McKenna (2004) observed the Internet has evolved into a useful tool to facilitate and maintain social communication and support networks.

A third objection to Internet-based research has to do with the lack of correspondence between its results and those derived from standard (paper and pencil) methods of data collection. However, across many areas of investigation, especially those involving survey or scale research, Internet and traditional forms of data collection yield parallel findings. Results obtained from online scales versus paper and pencil versions of the same scales have been found to be highly similar in survey research, although experimental designs, especially ones requiring attention to subtle stimuli, might be more vulnerable when undertaken on the Internet (Kraut et al., 1998). On the other hand, the more widely representative samples generally seen in Internet-based research must be figured into evaluations of the modality.

A fourth misconception is that participants completing a study online are not necessarily as motivated as those completing a study in person. The misconception that online participants tend to be unmotivated stems from the view that online studies are usually anonymous and could be completed by practically anyone, thus increasing the potential to answer questions carelessly. Because it is convenient to abandon a study in the middle if it appears uninteresting, dropout rate is indeed higher in Internet-based studies (Birnbaum, 2004). If a study proves boring to Internet participants, they are likely to vote with their feet – or mouse – and leave the study. Do not presume that a lackluster study failing to capture participant attention in the laboratory will somehow magically become exciting if it is instead implemented online.

Motivation to pay attention may be enhanced using financial incentives and lotteries, keeping time demands low (typically, 15 minutes or less), and offering participants automated feedback (e.g., a tailored personality profile resulting from responses reported on a personality inventory) immediately following completion of the online study. These fixes have introduced problems of their own. For example, if a monetary award is promised, researchers must be aware of the possibility of participants trying to take part in the experiment more than once to gain more money or to enhance their chances of winning a promised lottery. Many different solutions to the multiple submission problem have been used (see Musch & Reips, 2000), one of the most common of which prevents duplicate submissions from the same IP address. In addition, it is common practice to introduce various attention and manipulation checks in online research to help ensure data of high quality. The use of automated bots to answer surveys is a common fear among researchers, but rather simple measures can help offset the problem. Bots have trouble answering simple questions like, "What is your favorite flower (breed of dog, car, flavor of ice cream)?"

A major advantage of Internet-based research is that it removes the physical presence of the experimenter from the research setting. This does not imply that the online setting is free from artifacts stemming from experimenter expectancies and biases. Inadvertently providing clues to the research hypothesis should be avoided in online materials as well as in face-to-face

contexts. Some research results have suggested that data collected through online studies are largely equivalent to traditional studies, and therefore likely to result in greater power owing to the convenience of collecting larger samples. This is a practical consideration that should be considered (Beuckalaer & Lievens, 2009). However, even though the use of online participants may yield greater external validity due to the potential diversity of respondents, the generalizability of such findings is limited to people with Internet access who happen to volunteer, read, and understand the language of the instructions for your study. It may be that the Internet is a mixed blessing for researchers; however, in the best of all possible scenarios, using the Internet for online research may facilitate experimentation outside the laboratory, conveniently and efficiently. Such an approach allows us to draw on a much greater demographic range of participants, helping to enhance participant generalizability, and minimizing experimenter bias and demand.

Conclusion

Moving from laboratory to the field is for many experimentalists a daunting journey. The lab offers many advantages, including the potential to control many extraneous factors that can ruin an otherwise elegant test of important ideas. Researchers whose principal venue is outside the warm confines of the laboratory recognize this, and regret the loss of control the field sometimes imposes on their research. However, conducting research in either venue offers advantages the other cannot provide. We have recommended throughout a movement of researchers between laboratory and field, not merely to enjoy the unique benefits of each, but to provide a grounding of sometimes esoteric ideas in the tighter laboratory confines, while searching for developments whose widespread diffusion may benefit society and the quality of the theories being examined.

We see Internet as a kind of bridge between field and laboratory, a venue that is not as confined as the laboratory, and ideally not as free-range as social research conducted entirely in the field. As both venues offer many and varied advantages (and disadvantages), so too does research that takes advantage of Internet-based research. It offers the potential for greater generalizability than laboratory-based findings, but perhaps not as much as well-constructed field research. On the other hand, with sufficient ingenuity, the Internet can be an extraordinarily useful adjunct to social research. To ignore its possible benefits and advantages is foolhardy, just as it is to fail to recognize and avoid the many potential pitfalls the Internet user must surmount.

As we have maintained throughout this book, relying on a single method, a single source of participants, a single research context, the instantiation of a construct through ultra-dependence on a single exemplar, a single dependent measure for all our research needs, etc., is worse than foolhardy, it is self-deceptive. Every method has its strengths and weaknesses. In a spirit of multiple operationism, we fully endorse a multimethod approach in our research, and that includes use of multiple research contexts, laboratory, field, and now, the Internet, to test our hypotheses.

Notes

- 1 A useful sourcebook for considerations of reproducibility and replicability can be found in a book by the National Academies of Sciences, Engineering, and Medicine, 2019. The book is free to download.
- 2 In the United States alone, for example, nearly 90 million homes had Internet access in 2012, and of these, in households with annual incomes over \$50,000, an incredible 97% were wired.
- 3 A number of these are available, and they also allow for the random assignment of participants to experimental conditions. The most popular of these research-oriented programs are Qualtrics (www.qualtrics.com), Survey Monkey (www.surveymonkey.com), and Survey Gizmo (www.surveygizmo.com). Many others are available, and the list of possibilities seems to change often.

References

- Acerbi, A. (2022). From storytelling to facebook: Content biases when retelling or sharing a story. *Human Nature*, 33, 132–144.
- Alabastro, A., Rast, D. E., III, Lac, A., Hogg, M. A., & Crano, W. D. (2013). Intergroup bias and perceived similarity: Effects of successes and failures on support for in- and outgroup political leaders. *Group Processes & Intergroup Relations*, 16, 58–67.
- Amso, D., Cusack, R., Oakes, L., Tsuji, S., & Kirkham, N. (Eds.) (2022). *Empirical research at a distance: New methods for developmental science*. Lausanne: Frontiers Media SA. <https://doi.org/10.3389/978-2-88976-382-5>
- 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.
- Aspinwall, L., & Taylor, S. E. (1992). Modeling cognitive adaptation: A longitudinal investigation the impact of individual differences and coping on college adjustment and performance. *Journal of Personality and Social Psychology*, 63, 989–1003.
- Bargh, J. A., & McKenna, K. Y.A. (2004). The Internet and social life. *Annual Review of Psychology*, 55, 573–590.
- Beuckalaer, A. D., & Lievens, F. (2009). Measurement equivalence of paper-and pencil and Internet organizational surveys: A large scale examination in 16 countries. *Applied Psychology: An International Review*, 58, 336–361.
- Birnbaum, M. H. (2000a). Introduction to psychological experimentation on the internet. In M. H. Birnbaum (Ed.), *Psychological experiments on the internet* (pp. 3–34). San Diego, CA: Academic Press.
- Birnbaum, M. H. (Ed.) (2000b). *Psychological experiments on the internet*. San Diego, CA: Academic Press.
- Birnbaum, M. H. (2004). Human research and data collection via the Internet. *Annual Review of Psychology*, 55, 803–832
- Buchanan, T. (2000). Potential of the internet for personality research. In M. H. Birnbaum (Ed.), *Psychological experiments on the internet* (pp. 121–140). San Diego, CA: Academic Press.
- Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon’s Mechanical Turk: A new source for inexpensive, yet high-quality data? *Perspectives on Psychological Science*, 6, 3–5.
- Brady, J. (1958). Ulcers in “executive monkeys.” *Scientific American*, 199, 95–100.
- Busam, B., & Solomon-Moore, E. (2021). Public understanding of childhood obesity: Qualitative analysis of news articles and comments on facebook. *Health Communication*. <https://doi.org/10.1080/10410236.2021.1985859>
- Canary, H. E., Wellman, N., & Martinez, L. S. (2022). Covid-19, genetics, and risk: Content analysis of facebook posts early in the coronavirus pandemic. *Health Communication*. [doi.org.ccl.idm.oclc.org/10.1080/10410236.2022.2027639](https://doi.org/10.1080/10410236.2022.2027639)
- 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, 3156–2160.
- Cialdini, R. B., Borden, R. J., Thorne, A., Walker, M. R., Freeman, S., & Sloan, L. R. (1976). Basking in reflected glory: Three (football) field studies. *Journal of Personality and Social Psychology*, 34, 366–375.
- Cortez, V. L., & Bugental, D. B. (1995). Priming of perceived control in young children as a buffer against fear-inducing events. *Child Development*, 66(3), 687–696.
- Delle, F. A., Clayton, R. B., Jordan Jackson, F. F., & Lee, J. (2022). Facebook, Twitter, and Instagram: Simultaneously examining the association between three social networking sites and relationship stress and satisfaction. *Psychology of Popular Media*. <https://doi.org/10.1037/ppm00004153>
- De Simone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior*, 36, 171–181.
- Deters, F. G., & Mehl, M. R. (2013). Does posting Facebook status updates increase or decrease loneliness? An online social networking experiment. *Social Psychological and Personality Science*, 4, 579–586.
- Donaldson, C. D., Siegel, J. T., & Crano, W. D. (2016). Nonmedical use of prescription stimulants in college students: Attitudes, intentions, and vested interest. *Addictive Behaviors*, 53, 101–107.
- Evans, D. C., Garcia, D. J., Garcia, D. M., & Robert, R. S. (2003). In the privacy of their own homes. Using the internet to assess racial bias. *Personality and Social Psychology Bulletin*, 29, 273–284.

- Forgas, J. P. (Ed) (2000). *Feeling and thinking: The role of affect in social cognition*. New York: Cambridge University Press.
- Freedman, J. L., & Fraser, S. C. (1966). Compliance without pressure: The foot-in-the-door technique. *Journal of Personality and Social Psychology*, 4, 195–202.
- Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the web as good as the lab? Comparable performance from web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, 19, 847–857.
- Glass, D., & Singer, J. (1972). *Urban stress*. New York: Academic Press.
- Gonsalkorale, K., & von Hippel, W. (2011). Intergroup relations in the 21st century: Ingroup positivity and outgroup negativity among members of an Internet hate group. In R. Kramer, G. Leonardelli, & R. Livingston (Eds.), *Social cognition, social identity, and intergroup relations* (pp. 163–188). New York: Psychology Press.
- Gosling, S. D., & Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist*, 59, 93–104.
- Hales, A. H., & Williams, K. D. (2021). Social ostracism: Theoretical foundations and basic principles. In P. A. M. Van Lange, E. T. Higgins, & A. W. Kruglanski (Eds.), *Handbook of basic principles* (3rd Ed.) (pp. 337–349). New York, NY: Guilford Press.
- 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(1), 400–407.
- Isen, A. M., & Levin, P. F. (1972). The effect of feeling good on helping: Cookies and kindness. *Journal of Personality and Social Psychology*, 21, 384–388.
- Janis, I. L. (1958). *Psychological stress*. New York: Wiley.
- Johnson, J. E., & Leventhal, H. (1974). Effects of accurate expectations and behavioral instructions on reactions during a noxious medical examination. *Journal of Personality and Social Psychology*, 29, 710–718.
- Johnson, K. M., Iyer, R., Wojcik, S. P., Vaisey, S., Miles, A., Chu, V., & Graham, J. (2014). Ideology-specific patterns of moral indifference predict intentions not to vote. *Analyses of Social Issues and Public Policy*, 14, 61–77.
- Jones, S. (Ed.) (1999). *Doing internet research: Critical issues and methods for examining the net*. Thousand Oaks, CA: Sage.
- Katz, I. R., Resnick, S. G., Kaspro, W. J., Boden, M. T., Cherkasova, E., Fielstein, E. M., Trafton, J. A., & Hoff, R. A. (2020). Using patient-reported outcome measures for program evaluation: Design and findings on intention-to-treat outcomes from the veterans outcome assessment survey. *Psychiatry Research*, 291, 113226.
- Kiesler, S., & Sproull, L. S. (1986). Response effects in the electronic survey. *Public Opinion Quarterly*, 50, 402–413.
- Kochovska, S., Huang, C., Johnson, M. J., Agar, M. R., Fallon, M. T., Kaasa, S., Hussain, J. A., Portenoy, R. K., Higginson, I. J., & Currow, D. C. (2020). Intention-to-treat analyses for randomized controlled trials in hospice/palliative care: The case for analyses to be of people exposed to the intervention. *Journal of Pain and Symptom Management*, 59(3), 637–645.
- Krantz, J. H., Ballard, J., & Scher, J. (1997). Comparing the results of laboratory and world-wide-web samples on the determinants of female attractiveness. *Behavioral Research Methods, Instruments, and Computers*, 29, 264–269.
- Krantz, J. H., & Dalal, R. (2000). Validity of web-based psychological research. In M. H. Birnbaum (Ed.), *Psychological experiments on the internet* (pp. 35–60). San Diego, CA: Academic Press.
- Kraut, R., Patterson, V., Lundmark, M., Kiesler, S., Mukophadhyah, T., & Schlerlis, W. (1998). Internet paradox. A social technology that reduces social involvement and psychological well-being. *American Psychologist*, 53, 1017–1031.
- Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology*, 32, 311–328.
- Langer, E. J., & Rodin, J. (1976). The effects of choice and enhanced personal responsibility for the aged: A field experiment in an institutional setting. *Journal of Personality and Social Psychology*, 34, 191–198.
- Levine, M., Cassidy, C., Brazier, G., & Reicher, S. (2002). Self categorization and bystander non-intervention: Two experimental studies. *Journal of Applied Social Psychology*, 32, 1452–1463.
- Lutz, J. (2016). The validity of crowdsourcing data in studying anger and aggressive behavior: A comparison of online and laboratory data. *Social Psychology*, 47(1), 38–51.

- McCredie, M. N., & Morey, L. C. (2019). Who are the Turks? A characterization of MTurk workers using the personality assessment inventory. *Assessment*, 26(5), 759–766.
- McGraw, K. O., Tew, M. D., & Williams, J. E. (2000). The integrity of web-delivered experiments: Can you trust the data? *Psychological Science*, 11, 502–506.
- Medin, A. C., Astrup, H., Kåsin, B. M., & Andersen, L. F. (2015). Evaluation of a web-based food record for children using direct unobtrusive lunch observations: A validation study. *Journal of Medical Internet Research*, 17(12), 1–14.
- Mills, R. T., & Krantz, D. S. (1979). Information, choice, and reactions to stress: A field experiment in a blood bank with laboratory analogue. *Journal of Personality and Social Psychology*, 37, 608–620.
- Moghaddam, F. M., & Studer, C. (1998). *Illusions of control: Striving for control in our personal and professional lives*. Westport, CT: Praeger.
- Musch, J., & Reips, U. D. (2000). A brief history of web experimenting. In M. H. Birnbaum (Ed.), *Psychological experiments on the internet* (pp. 61–87). San Diego, CA: Academic Press.
- National Academies of Sciences, Engineering, and Medicine (2019). *Reproducibility and replicability in science*. Washington, DC: The National Academies Press.
- Necka, E. A., Cacioppo, S., Norman, G. J., & Cacioppo, J. T. (2016). Measuring the prevalence of problematic respondent behaviors among MTurk, campus, and community participants. *PLOS ONE*, 11(6), 1–19.
- Parker, S. D., Brewer, M. B., & Spencer, J. R. (1980). Natural disaster, perceived control, and attributions to fate. *Personality and Social Psychology Bulletin*, 6, 454–459.
- Piliavin, I. M., Rodin, J., & Piliavin, J. A. (1969). Good Samaritanism: An underground phenomenon? *Journal of Personality and Social Psychology*, 13, 289–299.
- Quist, R. M., & Crano, W. D. (2003). Assumed policy similarity and voter preference. *Journal of Social Psychology*, 143, 149–162.
- Reips, U. D. (2000). The web experiment method: Advantages, disadvantages, and solutions. In M. H. Birnbaum (Ed.), *Psychological experiments on the internet* (pp. 89–117). San Diego, CA: Academic Press.
- Rudert, S. C., Greifeneder, R., & Williams, K. D. (2019). *Current directions in ostracism, social exclusion, and rejection research*. New York, NY: Routledge/Taylor & Francis Group.
- Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, 51, 515–530.
- Sequeira, L., Perrotta, S., LaGrassa, J., Merikangas, K., Kreindler, D., Kundur, D., Courtney, D., Szatmari, P., Battaglia, M., & Strauss, J. (2020). Mobile and wearable technology for monitoring depressive symptoms in children and adolescents: A scoping review. *Journal of Affective Disorders*, 265, 314–324.
- Sigelman, L. (1986). Basking in reflected glory revisited: An attempt at replication. *Social Psychology Quarterly*, 49, 90–92.
- Simmons, A., McGatlin, K., & Lustig, C. (2022). How well do online, self-administered measures correspond to in-lab assessments? A preliminary examination of three measures in healthy older adults. *Neuropsychology*. <https://doi.org/10.1037/neu0000819> 10.1037/neu0000819.supp (Supplemental)
- Skitka, L. J., & Sargis, E. (2006). The Internet as psychological laboratory. *Annual Review of Psychology*, 57, 529–555.
- Smith, R. A., & Boster, F. J. (2009). Understanding the influence of others on perceptions of a message's advocacy: Testing a two-step model. *Communication Monographs*, 76(3), 333–350.
- Stoll, C. (1995). *Silicon snake oil*. New York: Doubleday.
- Sun, Y., & Lu, F. (2022). How misinformation and rebuttals in online comments affect people's intention to receive COVID-19 vaccines: The roles of psychological reactance and misperceptions. *Journalism & Mass Communication Quarterly*. <https://doi.org/10.77/107769902210846>
- Taylor, S. E., & Aspinwall, L. G. (1996). Mediating and moderating processes in psychological stress: Appraisal, coping, resistance, and vulnerability. In H. B. Kaplan (Ed.), *Psychological stress: Perspectives on structure, theory, life-course, and methods* (pp. 71–110). Cambridge, MA: Academic Press.
- Thompson, S. C., & Collins, M. A. (1995). Applications of perceived control to cancer: An overview of theory and measurement. *Journal of Psychosocial Oncology*, 13, 11–26.
- Turkle, S. (1996). Virtuality and its discontents: Searching for community in cyberspace. *The American Prospect*, 24, 50–57.
- Underwood, B., Berenson, J., Berenson, R., Cheng, K., Wilson, D., Kulik, J., Moore, B., & Wenzel, G. (1977). Attention, negative affect, and altruism: An ecological validation. *Personality and Social Psychology Bulletin*, 3, 54–58.

- Vancleef, L. M. G., & Peters, M. L. (2011). The influence of perceived control and self-efficacy on the sensory evaluation of experimentally induced pain. *Journal of Behavior Therapy and Experimental Psychiatry*, 42(4), 511–517.
- Walther, J. B., DeAndrea, D., Kim, J., & Anthony, J. (2010). The influence of online comments on perceptions of anti-marijuana public service announcements on YouTube. *Human Communication Research*, 36, 469–492.
- Washington, P., Kalantarian, H., Tariq, Q., Schwartz, J., Dunlap, K., Chrisman, B., Varma, M., Ning, M., Kline, A., Stockham, N., Paskov, K., Voss, C., Haber, N., & Wall, D. P. (2019). Validity of online screening for autism: Crowdsourcing study comparing paid and unpaid diagnostic tasks. *Journal of Medical Internet Research*, 21(5), e13668.
- Webb, E. J., Campbell, D. T., Schwartz, R. D., Sechrest, L., & Grove, J. (1981). *Nonreactive measures in the social sciences*. Boston, MA: Houghton-Mifflin.
- Williams, K. D. (1997). Social ostracism. In R. Kowalski (Ed.), *Aversive interpersonal behaviors* (pp. 133–170). New York, NY: Plenum.
- Williams, K. D., Cheung, C., & Choi, W. (2000). Cyberostracism: Effects of being ignored over the internet. *Journal of Personality and Social Psychology*, 79, 748–762.
- Williams, K. D., & Zadro, L. (2005). Ostracism: On being ignored, excluded, and rejected. In M. R. Leary (Ed.), *Interpersonal rejection* (pp. 21–53). New York, NY: Oxford University Press.
- Wilson, R., Gosling, S., & Graham, L. (2012). A review of Facebook research in the social sciences. *Perspectives on Psychological Science*, 7, 203–220.

9

QUASI-EXPERIMENTS AND APPLIED RESEARCH

In the preceding chapters we distinguished experimental from correlational research, in which the investigator's role is largely that of an observer. In a real sense, all variables in correlational studies are dependent (or response) variables, and the researcher's job is to assess their natural variation and discern the patterns and relations among them. In experiments, the researcher actively intervenes in the normal relational pattern by systematically controlling variation in the independent variable (or variables) to assess the causal impact of this variation on outcome variables. Controlled variation of the (presumed) causal variable and random assignment of subjects to research conditions are the necessary hallmarks of true experiments, which are the backbone of internally valid cause-effect analyses.

In many research contexts, the distinction between experimental and correlational studies may not be clear. For example, in our discussion of field experiments, we mentioned studies in which the researcher *selected* rather than *created* the levels of the independent variable. Such studies preserve the logic of experimental design but lack the degree of experimenter control that characterizes causal analysis. By the same token, some correlational studies are conducted in the context of interventions into a given social situation (e.g., studies that investigate acceptance of flood-abatement information by different groups who are at different levels of flood risk (e.g., De Dominicis et al., 2014, 2021), thus mixing aspects of experimental and correlational designs).

Intermediate to true experiments and pure correlational research are quasi-experimental studies in which some systematic variation in exposure to different conditions has occurred between groups, but random assignment of participants to these groups was not, or could not be employed. This could occur because some groups experienced an intervention or naturally occurring "treatment," whereas other groups did not, and the researcher had no control over who did or did not receive it. Some participants may have been exposed to a treatment while others were not because of a non-random process. These situations bear some resemblance to the pretest-posttest control group experimental design, in that a treatment has been introduced at a specifiable point to one group and not to another, except that the researcher had no control over who was placed in these groups. Differences observed between measures taken before the introduction of the treatment in a single group and those occurring after its introduction mimic true experimental designs, but the structure of such studies lacks a critical element necessary for internal validity. In the absence of random assignment of participants to experimental

conditions, it is impossible to separate effects caused by the introduction of the treatment from effects attributable to prior group differences, historical events, and other threats to internal validity. Although quasi-experiments often have fewer threats to internal validity than correlational studies, they are generally (but not always) more susceptible to such threats than randomized experiments.

Quasi-Experimental Methods in Applied Contexts

The social research settings in which non-randomly assigned treatment conditions are commonly encountered is in **evaluation research**, where the goal of research is to assess the impact or effectiveness of some social program or policy. When feasible, evaluation researchers will employ true experimental designs, as they provide the surest route to causal inference about program effects. As we demonstrated in Chapter 8, randomization is possible in many field settings, and examples of randomized experiments in programs of research and evaluation are available (Boruch, 2012; Green et al., 2011), even with programs involving large-scale implementation (e.g., “negative income tax” programs, Kershaw & Fair, 1976; Levine et al., 2005), manager’s mindfulness training and well-being (Vonderlin et al., 2021), government funding of nonprofits (Jilke et al., 2019), voting behavior (Arceneaux, 2005), the effects of experimental innovative electric rate variations on energy conservation (Crano & Messé, 1985), and a large study of an HIV/AIDS prevention program in Maryland, Georgia, and New Jersey (O’Leary et al., 1998).

Although good arguments can be made for the use of random assignment through lottery-based assignment as a way of allocating a scarce resource or service in the interests of fairness (e.g., Krueger & Zhu, 2004; Salovey & Williams-Piehota, 2004), randomization often is not politically or practically feasible. Sometimes it is impossible to control who will make use of available services or programs (e.g., who will choose to attend a free medical clinic). At other times, programs can be delivered selectively, but the selection decision is outside the researcher’s control, based on non-random factors such as perceived need (the Head Start or negative income programs) or merit (Gifted and Talented school programs, academic or sports scholarships), or political considerations. When this occurs, researchers often seek quasi-experimental alternatives to sort out likely treatment effects from other sources of variation.

Quasi-experimental designs maintain many of the features of true experiments without the advantages of random assignment. The absence of random assignment and the presence of some treatment variation between groups are defining features of quasi-experiments and require researchers to seek ways to offset the problems of causal interpretation that consequently arise (Campbell & Stanley, 1963; Shadish et al., 2002). For the remainder of this chapter, we will be concerned with various quasi-experimental designs and the issues with their use in research.

Regression Artifacts and Assessment of Change

Because programs often are introduced into ongoing social systems to test or improve some aspect of that system, the ultimate question for researchers is whether the system or the persons in it have *changed* over time because of the program. To understand how the nature of research design affects our ability to assess change meaningfully, we must consider the critical issue of *regression toward the mean* as a potential artifact in the measurement of change, which was introduced in Chapter 2 as a potential threat to internal validity.

The origin of the term *regression toward the mean* provides some insight into this effect. Sir Francis Galton (1885) introduced the issue in his paper, “Regression towards mediocrity in hereditary stature,” in which he reported the results of his study of the relation between the height of parents

and their adult offspring. One finding of this study was that the adult children of very tall parents were on average not quite as tall as their parents, whereas the adult children of very short parents were on average not quite as short. Usually, the heights of offspring of extreme individuals were closer to the overall population average than that of their parents – the heights of the offspring “regressed” toward the mean of the population from which they were drawn. The trend observed by Galton is often termed the “regression fallacy,” because such results often are misinterpreted as indicating a long-term tendency toward mediocrity. It does not. The implication of this erroneous view is that across generations, variations in parent-child heights become smaller and smaller as the concentration of individuals who are farther away from the mean height of the population diminishes. In fact, the variation does not necessarily change from generation to generation, because even though the offspring of extreme parents tend to be closer to the mean height than their parents were, the offspring of more average parents are equally likely to vary away from the population mean, closer to either end of the two extremes.

The apparent shrinkage is an artifact of the way respondents were selected. Phrased differently, very tall parents still tend to have tall offspring (relative to the average of the population), but the offspring tend not to be as extremely tall as the parents. The reverse applies for very short parents and their short offspring. The movement toward the mean in a subsequent measurement is an artifact of the initial selection of extreme cases.

Regression toward the mean is an *inevitable* consequence of examining the association between scores on imperfectly related measures. Whenever the correlation between two measures (like parental height and offspring height) is less than 1.00, there will be some nonsystematic or random deviation between scores on the first variable and corresponding scores on the second. If the first set of scores are selected for extremity (i.e., to represent the highest or lowest values in the distribution), there is bound to be a bias in what direction subsequent scores using the same measuring instrument will vary. For the tallest parents in a population, any differences between their heights and those of their offspring will usually indicate that their offspring are somewhat shorter, simply because there isn’t much room for variation (ceiling effects) in the other direction. Similarly, deviations from the heights of extremely short parents most often will be in the direction of increased height because of a similar selection bias (floor effects).

It is important to distinguish between findings that result from artifacts and those that reflect real effects. An **artifact** is an artificial or spurious finding, a “pseudo-effect” that results inevitably from the properties of the measuring instrument or from the method of data collection employed. In Galton’s example, regression does not reflect some genetic defect on the part of extreme parents that results in mediocre offspring. Extremely tall parents generally produce tall children and extremely intelligent parents usually have highly intelligent children. However, the relationship between characteristics of parents and offspring is not perfect. Because of this, the selection of parents based on an extreme characteristic *biases* the direction of differences between parents and children in a way that has nothing to do with the laws of genetics.

Regression artifact and reliability. The regression artifact afflicts psychological research most frequently because of measurement unreliability, which is responsible for imperfect correlations between repeated tests on the same measure. Test reliability has been referred to in previous chapters, but a detailed consideration is useful here to clarify the role of unreliability in the occurrence of regression effects.

The basic elements of the measurement theory model of test-retest reliability are presented in Table 9.1. Two sets of test scores (pretest and posttest) are obtained on a measure administered twice to the same sample of 20 participants. The data in this table indicate that each test score is assumed theoretically to consist of two components – true score and error (see Chapter 3). The true score component represents stable characteristics of those who have taken the measure.

TABLE 9.1 Random error and test-retest reliability, with response and two error scores for each respondent, and resulting Test 1 and Test 2 final scores, respectively

<i>Respondent number</i>	<i>True score</i>	<i>Test 1 Random error</i>	<i>Test 1 Score</i>	<i>Test 2 Random error</i>	<i>Test 2 Score</i>
101	95	−5	90	+1	96
142	93	+2	95	−3	90
103	92	−6	86	0	92
210	90	+8	98	−7	83
221	87	+1	88	+1	88
043	85	−5	80	+6	91
110	85	+5	90	+3	88
042	80	−3	77	+1	81
303	78	+6	84	−7	71
370	75	+9	84	+6	81
222	75	−7	68	+4	79
260	74	−5	69	+1	75
157	73	+6	79	−8	65
192	70	−2	68	−4	66
377	68	−3	65	−2	66
298	65	−4	61	+3	68
119	63	+3	66	+5	68
327	60	−2	58	+4	64
253	58	+5	63	−3	55
399	55	−3	52	−1	54
Sum = 1,521		0	1,521	0	1,521
Mean = 76.05		0	76.05	0	76.05
SD = 12.21			13.23		12.60

Source: On the bottom rows are the average scores of the total group, with the total scores' standard deviations, to illustrate the effects of error on scores. In this case, the average score is the same, because the error scores in each test administration cancel.

We assume the true score for any individual being measured does not change between test administrations unless some fundamental change has occurred in the individual's belief system, or some contextual upheaval has caused a serious reconsideration of formerly normal beliefs. Neither of these has occurred, and thus, in Table 9.1 (which illustrates test-retest relationships under no-change conditions), each of the 20 hypothetical individuals (whose ID Respondent Number ranges from 101 to 399 in the table) is represented by their underlying true score, and two error scores which contribute to the obtained score on Test 1 and a later Test 2. This is an idealized case, where the error scores within each test cancel. This is not often the case, and so some differences between the test scores might be expected. However, the Test 1 to Test 2 totals should not differ significantly if no events that might affect the true score intervene.

In contrast to the stability of true score variation, the error component of test scores represents all the temporary, random factors that influence test responses at a particular time point. The most important assumption of testing theory is that these chance factors are random. That is, some individuals' scores are artificially raised by these chance factors, and others' scores are lowered, so that across many individuals error effects cancel. This characteristic of error scores is illustrated in Column 3 of Table 9.1, where the algebraic sum of the 20 error scores (which represent the extent and direction of random influences on each respondent's test score) is equal to 0. The test score obtained by combining the true score component of Column 2 with the

error score from Column 3 for each individual is given in the fourth column. It represents the observed score, that is, the score an individual would receive on the measure.

Because the error scores cancel each other, the sum and mean of the obtained test scores are the same as the corresponding values for the true scores. However, the variability of obtained scores is different from that of the corresponding underlying true scores because of the effects of random error. As shown, the standard deviations (SDs) of the error-free (Column 2) and the observed score of Column 3 are different, and this enlargement of the standard deviation is directly attributable to the effect of random error.

If error scores are truly random, then the extraneous factors that determined the direction and extent of error on one testing should *not* be the same in a second testing of the same group of individuals, assuming a reasonably large sample. For any individual, the random error component of the score on a first testing should be unrelated to the random error score on the second testing. Otherwise, a directional bias affecting scores at both testing administrations would suggest that the measurements were affected by systematic error. In Column 4 of Table 9.1, a different set of random error scores is shown, affecting participants' responses at the second administration. These scores represent the random influences on responses for each respondent at the time of retesting. Although the algebraic sum of these Time 2 error scores is equal to 0, as on the first testing, the pattern of errors across individuals is different from that of the first testing. The sum of true score and error for the second testing results in the obtained score values for Test 2 are recorded in the final column of Table 9.1.

The combined effect of unrelated error components introduces discrepancies in test results for the same individuals across different testing occasions in repeated measure designs. For any individual, a test score obtained from a second measurement is expected to deviate somewhat from that person's score on the same instrument at the first testing. The extent of deviation in scores between tests reflects the degree of test-retest unreliability of the measurement instrument, along with the effect of any systematic effects. The more unreliable the test-retest association, the greater the error, and as such, the greater will be the deviations in scores from the first administration to the next. The degree of *similarity* in patterns of scores across different administrations of the same test is termed the test-retest reliability. Because reliability involves the relationship between two sets of test scores, it is most often assessed with the Pearson correlation (see Chapter 10). The reliability coefficient is the value of the Pearson correlation between the test results of a set of individuals at time 1 with the results for these *same individuals* on the *same test* at time 2. For the data in Table 9.1, the value of the reliability coefficient is equal to .82, which indicates that those who scored high at time 1 also tended to score high at time 2, and those who scored low at time 1 tended correspondingly to score low at time 2.

As discussed in Chapter 10, the squared value of the correlation coefficient (r^2 , the coefficient of determination) measures the extent of common variation between two measurements. In the case of reliability coefficients, the correlation value represents the proportion of true score variation in the measuring instrument, or the extent that obtained scores at both assessments reflect true score values rather than chance error. The lower the reliability correlation, the greater the proportional effect of error (defined as $\epsilon = 1 - r^2$), and the more that random extraneous factors influence the value of obtained test scores. A perfectly reliable instrument ($r = 1.00$) would reflect only true score variation, and thus produce no regression effects because there would be no discrepancies between scores on first and second test administrations in the absence of an intervening event. For our hypothetical data, the value of r indicates that 67% ($.82^2$) of the obtained variation in test scores can be attributed to true score differences, and the remaining 33% is due to random fluctuation. Unreliability in tests is responsible for the internal validity threat of regression toward the mean in test-retest research studies. As can be seen, although the

refinement of most tests is aimed at achieving reliability values of .70 or better, the remaining error component in this case ($\epsilon = 1.0 - .70^2$) could produce considerable regression between tests of groups selected because of their extremity.

Statistical control of regression artifacts. A method of controlling for the effects of regression toward the mean in pretest-posttest research designs involves statistically removing the expected regression toward the mean effect in computing difference scores. This is accomplished by using multiple regression analysis and comparing each individual's actual posttest score (Y) with a "regressed" score, that is, a predicted posttest score (Y') based on the expected deviation from his or her pretest score. Regressed scores are obtained by multiplying the pretest score in standardized score form by the value of the reliability coefficient (i.e., the test-retest correlation). Across all respondents, the average obtained posttest score should not differ from the average of the regressed scores if no true score change has occurred. If a significant difference between predicted and obtained scores does appear, it indicates that some change has occurred above and beyond that expected, based on regression toward the mean.

A difficulty with the statistical adjustment of regression effects is that it requires an accurate assessment of test-retest reliability in the absence of any real change. Computed values of the correlation between pretest and posttest scores suffer from the same sampling fluctuation as any other statistic, and in addition, the appropriate time lag necessary for useful test-retest reliability assessment is difficult to determine. Testing must be distant enough in time to assure that for each respondent, the sources of error on the second testing (such as mood, fatigue, misinterpretation of questions, guessing behavior) are unrelated to those of the first test session. The test-retest time lag also should be appropriate to provide a base estimate of normal test score fluctuation against which to compare treatment effects. It is rare that such an estimate of reliability is readily available.

Regression toward the Mean and Matching: Comparison Group Designs

Considering how regression toward the mean influences change scores from a single population with initially extreme test values are understandable from the standpoint of classic test theory. Because this effect is well known, simple pretest-posttest designs comparing groups chosen based on extreme pretest scores usually are avoided in program evaluation or other research contexts. However, the effects of regression artifacts can appear subtly, whenever research involves two or more respondent groups that are not randomly assigned to treatment and comparison conditions. This possibility arises, for example, when individuals participating in a social program have been selected non-randomly. When evaluating program effects, we often find that those receiving the planned intervention differ from comparison group participants in some systematic way on factors that may affect the outcome measures, even prior to the program intervention. This initial nonequivalence attributed to selection biases between groups makes it difficult to interpret differences between them.

Attempts to compensate for pretest inequality between the groups are made sometimes through post hoc participant "matching," a procedure used despite a history of frequent warnings against it (Campbell & Stanley, 1963; Hovland et al., 1949; McNemar, 1940; Thorndike, 1942). Post hoc matching involves making post-treatment comparisons between scores of the members of the two or more groups who attained similar scores on a pretest (or some other measure used to facilitate matching). In effect, this procedure amounts to an attempt to apply the blocking design of experimental research (discussed in Chapter 5), but the blocking of post hoc matching is accomplished *after* assignment to treatments, rather than before, and without initial random assignment of experimental and control groups. In this case, the different groups

may regress to different means because of a regression artifact and produce spurious outcomes that are solely a function of error, the cause of the artifact.

A slight variation of this matching technique occurs when an experimental treatment group is composed by meeting some special criteria and then a “control” group is constituted later by selecting a group of participants who “match” the predetermined experimental respondents on some specified pretest value(s). In this case, interpretation of the results of such matched group designs may be confounded by the potential effects of regression toward the mean.

Campbell and Erlebacher (1970) provided a detailed discussion of this type of regression artifact as it affects the interpretation of research on compensatory education programs. In this study, children attending a Project Head Start program were matched with a matched control group from a group of children whose families exceeded the low-income limit the program required. Often, special education programs are reserved for particularly disadvantaged populations. Students are not assigned to them randomly. Any attempts to evaluate their effectiveness through *ex post facto* comparison with a “matched” control group selected from a different population (e.g., a more affluent population than the economically disadvantaged respondents receiving the experimental intervention) introduces the danger of a regression bias that operates against the apparent value of the treatment, because matches between students from the two groups may be chosen from different ends of their respective scoring distributions. For example, matching students from different score distributions of a school readiness test may pair children from low-income families with those from a considerably higher income strata. This pairing requires that students from the top end of Head Start score distribution are paired with students from the lower end of the more well-to-do distribution, to allow the research to match them on scores. Students’ post-intervention scores will regress in opposite directions on the posttest, resulting in the apparent ineffectiveness of the program when, in fact, a “nonsignificant difference” between the groups on posttest would indicate a positive Head Start outcome, with the degree of positivity a function of the reliability of the posttest. A posttest of lower reliability would be expected to show a greater regression effect – overcoming such an effect would be a powerful signal of the success of the program (see Figure 9.1).

Post hoc matching based on extreme score pretreatment variable, or combination of variables (e.g., age, socioeconomic status, intelligence, personality scores), is subject to the same regression phenomenon if the matching variables are imperfect (i.e., not perfectly reliable), or imperfectly related to the post-treatment dependent variable measure. Such is the case, for example, with research that attempts to assess racial differences by comparing groups differing in race but “matched” on socioeconomic variables. In a society where race is still inextricably confounded with multiple cultural factors, such *post hoc* matching can only produce groups that are subject to differential regression effects, resulting ultimately in uninterpretable results.

Propensity Score Matching

In the standard matching design, it would be cumbersome for the researcher to match pairs of respondents on more than one or two criteria (say, sex and age). Suppose a program evaluator was asked to assess the efficacy of a campaign to promote walking as a form of exercise, with instructional brochures mailed to people living in a targeted community (the experimental group). Because the campaign targeted all residents in that neighborhood, a comparison community would be needed to determine if the campaign was effective in increasing exercising relative to a community that did not receive the mailing (the control group). After the campaign, people from each community might be sent a survey asking them about how many minutes they spent walking per day. Because it is possible that participants from disparate communities tend to differ

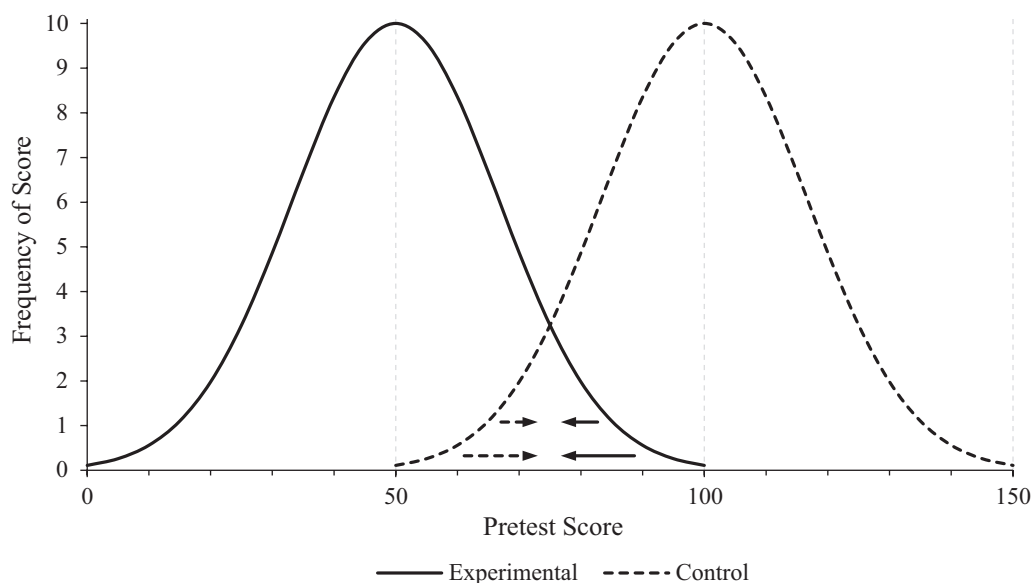


FIGURE 9.1 Illustration of the dangers of matching two groups from different score distributions. The regression artifact tends to lower the score of the experimental group (solid line) while it tends to increase the score of the control group (dashed line), because the matched group, which is drawn from the overlap between the two distributions, pairs the highest scoring participants from the intervention group and the lowest scoring participants from the control group. In both instances, the regression artifact tends to pull the two groups' scores toward their respective means, as shown by the arrows in the overlap between the two distribution.

on various characteristics (e.g., age, gender, SES, race, etc.), it is necessary to control for these extraneous covariates so that they do not serve as rival explanations to account for differences in time spent walking. For example, if men were more likely to engage in physical activity, and one community had many more men than the other, this would confound outcome differences.

In the standard matching design, respondents from the different communities are paired on the measured covariates of race, gender, etc., which takes the place of a sorting process that matches respondents so that a one-to-one correspondence is attained using multiple sorting variables. This would be prohibitively daunting as the number matching variables increased. A way of overcoming the limitations of matching is **propensity score matching** (Rosenbaum & Rubin, 1983), which uses statistical procedures to match respondents on as many covariates (i.e., matching variables) as specified by the researcher to determine differences between comparison groups. Results from propensity score matching reveal differences on an outcome measure between groups after relevant covariates have been accounted for.

A **propensity score** for a participant represents the conditional probability of membership in one group (e.g., the experimental group) over another (e.g., the control group), given the pattern of the person's status on the various covariates. Propensity scores are based on the entire set of covariate measures, and are used to adjust and control for these covariates statistically, so groups initially comparable on the covariate variables can be assembled. Operationally, responses on the covariates are used to compute a propensity score for each respondent. In a randomized two-group design, the propensity score of every participant is expected to be .50

(a 50% chance of being assigned to either group). In a nonexperimental or quasi-experimental design, because of non-randomized participant selection into groups, as in our exercise campaign scenario, propensity scores probably vary across participants, and matches on the scores between the groups being compared is the reason for the approach. Among the matched groups, the researcher has to find a twin for each respondent in the other group, or at least a person with similar, if not identical scores on variables thought to be important impingements on the outcome (in our case, age, gender, etc.).

The underlying logic is that if participants from different groups have the same propensity score, created by the statistical combination of many variables, then it is possible to deduce what each person's posttest score *would be* if they were in the other group, even if they were not. The estimated or hypothetical outcome score on the group or condition a participant was not a member of is known as the *counterfactual score*. After matching, the counterfactual posttest scores are compared. If those in the treatment group are different from respondents from the matched group, the difference is attributed to the treatment. The major assumption of propensity score matching is that it is possible to deduce mathematically what someone's hypothetical or counterfactual score would have been in the other group, based on their patterns of responses to the measured covariates.

Using propensity scores, participants across groups are made statistically equivalent *on these initial variables*, so that any posttest observations should rule out these characteristics as confounds. However, this approach is not without problems (see Crano et al., 2015). Although the technique allows for the matching on many pretest covariates, the extent and quality of the control is determined by the number of extraneous variables that were identified, relevant, and measured in the study. For the technique to be most effective, the researcher must consider and assess as many of the potentially important extraneous variables that can be identified, and on which groups may differ initially. Some highly stringent propensity score studies have controlled for more than 100 pretest covariates. Owing to the number of measures on which participants across groups can potentially be matched, and the many combinations of patterns of covariate responses, propensity score approaches often necessitate very large sample sizes.

Propensity scoring is one of the most methodologically acceptable approaches when contemplating a matching design. However, even after controlling for a wide assortment of variables, uncertainty always remains that there are lurking variables the researcher failed to measure, and therefore pretest equivalence was not attained. Propensity scoring is advantageous relative to other statistical techniques that control for covariates insofar as the assumptions about the data are more relaxed in propensity analysis. For example, propensity scoring does not require the assumption of linearity between each covariate and the outcome, as required in analysis of covariance. Even so, randomization to groups remains the most methodologically rigorous and satisfactory method to ensure the initial comparability of groups that are to be compared. Propensity score matching controls for only those extraneous variables that are measured and entered as part of the propensity score, whereas randomized designs control for both measured and unmeasured extraneous variables.

Quasi-Experimental Designs

The logic of quasi-experimentation requires consideration of design principles that recognize and minimize rival alternative hypotheses, and although the principles do not mitigate these rivals as surely as true experiments, their control helps enhance the plausibility of our interpretations (Shadish et al., 2002). The first of these principles requires the researcher to identify likely threats to internal validity as part of the research planning process. In some situations,

not all potential threats listed earlier (Chapter 2) are plausible, and so can be dismissed. Others, however, may support explanations of results that are different from those of the researcher, and must be dealt with to the extent possible. This leads to the second principle that prioritizes the importance of design over statistical control. In our view, statistical approaches to offset design weaknesses are always the second alternative. As with true experiments, designing studies that render rival explanations implausible is by far the preferable route. The third principle is concerned with “pattern matching,” which involves consideration of the match between the pattern of data obtained with the pattern expected based on the hypotheses. If the theorized pattern involves complex differences over time, and the obtained data match expectations, it lends credibility to the tenets of the theory. Even interaction effects can prove confirmatory if rival alternative hypotheses are focused only on main effects, while the theory predicts the interaction. Enhancements suggested for various quasi-experimental designs to be discussed were developed in accord with the three principles to offset the internal validity threats that impinge on different experimental designs.

Nonequivalent Groups Designs

The nonequivalent groups designs resemble those found in true experimental designs, with the critical difference being the lack of random assignment. This does not mean the quasi-experimental versions are not worth conducting. On the contrary, with proper enhancement the design can be informative, even though not allowing for confident causal inference. As shown on the following pages, nonequivalent groups designs vary considerably, but their common problem stems from the likely nonequivalence of the participant samples being compared.

Single-group, treatment-posttest design. The most primitive of the nonequivalent group designs uses only a single group, and sometimes does not use even a pretest. Diagrammatically Figure 9.2 involves a participant group that is pretested (or not), exposed to an intervention or treatment, and its effects measured.

In most cases, such designs provide weak evidence of causation, but these designs can encourage further, more systematic research. For example, suppose a math teacher of first year high school students has a hypothesis that teaching certain skills in her introductory algebra course would have positive effects on the students’ ability to learn the calculus in their fourth year. She tries this out in a cadre of first year students and waits three years for the results on the students’ introductory calculus scores in their senior year of high school. The researcher finds the scores of this cadre of students on their pre-calculus test are on average 30 points higher on a 100-point measure than any previous group in the school’s history (see Figure 9.3).

Although the researcher cannot claim her teaching innovation *caused* this miraculous result, there is no doubt that she would be encouraged to follow-up her quasi-experiment with more systematic, experimental research. The primitive design she used did not support claims of causation, but it stimulated further investigation.

Less primitive nonequivalent groups designs. Suppose the math teacher followed her initial apparent success with another quasi-experiment by teaching her innovating technique to her early morning class, but not to a class she taught later in the day. Both groups of students were in their first year of high school. The diagram of this design is shown in Figure 9.4.

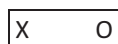


FIGURE 9.2 One-group pretest-posttest quasi-experimental design, where X = the experimental intervention, and O = the outcome (dependent) measure.

X	O ₁	O ₂

FIGURE 9.3 One-group pretest–posttest quasi-experimental design, with immediate O₁ and 3-year delayed outcome measure O₂, where X = the experimental intervention, and O₁ and O₂ = the outcome (dependent) measures.

O ₁	X	O ₂
O ₁		O ₂

FIGURE 9.4 A two-group pretest (O₁)–posttest (O₂) control group quasi-experimental design. The design is quasi-experimental because the participants were not assigned randomly to the experimental or control group.

This design improves the original one but remains quasi-experimental owing to the non-random selection of subjects into the two conditions. However, if the two groups had been assigned randomly to classes, then the study becomes an experiment, with two rival alternative hypotheses: (1) Although the students were assigned to classes randomly, the treatment was applied only in the morning class. This temporal difference, too, distinguished the two experimental groups, and represents a confound of time of application with the application itself.¹ (2) In addition, the children might have experienced a different educational experience owing to the classrooms in which they were taught. Although the analysis of this possible effect can be assessed (see ICC in Chapter 11), it adds some complexity to the analysis. Although we cannot definitively rule out these rival alternative explanations, the study certainly would encourage more refined experimentation. This process represents the accretive nature of scientific research, where one study rarely answers all questions, but the better studies often suggest ways to gain greater insight into a phenomenon.

Improving nonequivalent groups designs. Two established methods of improving the interpretability of quasi-experimental designs involve adding groups to address specific rival alternative hypotheses and matching obtained results with hypothesized patterns expected based on prior theory or earlier research. Adding groups sometimes can produce persuasive evidence of a hypothesized alternative’s plausibility. For example, suppose we have developed a method to teach multiplication to children in their second year of elementary school. We receive permission to test children from two difference schools. We choose School A’s students to receive the treatment, and School B’s students to act as the control group. The study is quasi-experimental, because we could not assign students randomly to conditions. We use a pretest–posttest control nonequivalent groups quasi-experimental design. Analysis of the data shows a clear treatment effect (see Figure 9.5). The intervention worked! But wait, although our groups differed on the posttest, they started with different scores. The experimental subjects were different (perhaps significantly different) from the controls on the pretest, and while the posttest difference was greater than the pretest difference it is possible the experimental participants’ multiplication readiness was developing faster than that of children in the control group. This rival alternative “developmental differences” hypothesis accounts for the achievement gap that grew as children progressed over the school year.

We can address this rival alternative hypothesis by adding an additional pretest to the design (Figure 9.6). If the relations of the first and second pretests are *parallel* when comparing the two groups, there is no evidence to support the development difference hypothesis. Both groups are progressing at similar rates; however, when the more advanced group receives the intervention,

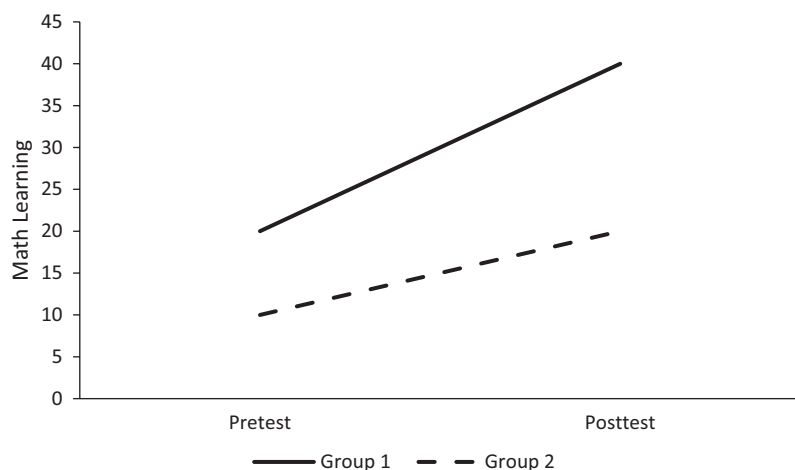


FIGURE 9.5 Results of a quasi-experimental pretest-posttest control group study of a hypothetical investigation of a technique to improve fourth-grade students' math learning. Group 1 is the intervention group. Group 2 is the control group.

its rate of growth far exceeds that of the control group. It is difficult to attribute this to a developmental jump at precisely the time at which the intervention was applied. On the other hand, the effectiveness of the intervention seems a reasonable, if tentative hypothesis.

By adding the “pre-pretest” to the design, we have rendered the alternative hypothesis of differential development rates implausible. Adjustments like the one here, simply adding an additional pretest, often can provide information useful in addressing a strong rival hypothesis. Such adjustments, if anticipated, can enhance the interpretability of quasi-experiments.

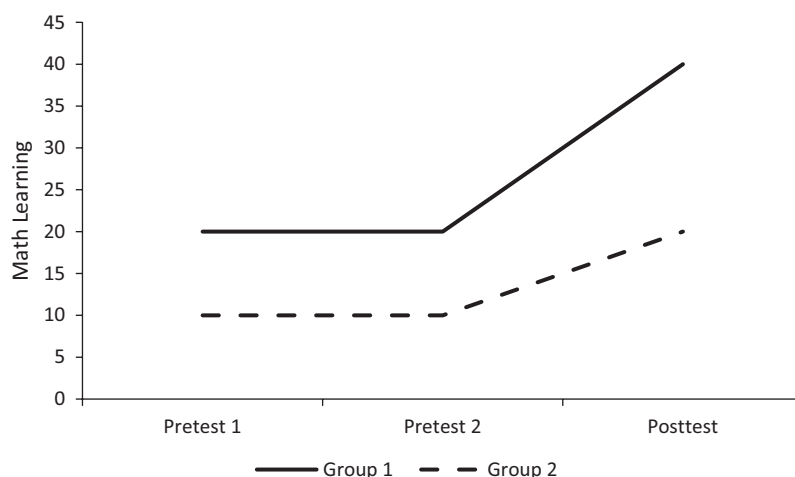


FIGURE 9.6 Results of the same quasi-experimental pretest-posttest control group study of a hypothetical investigation of a technique to improve fourth-grade students' math learning. Group 1 is the intervention group. Group 2 is the control group. In this design, multiple pretests are used to investigate the plausibility of the rival alternative hypothesis that maturation-rate differences explain the results. As shown by the parallel results between the two pretests, both groups appear to have been maturing at approximately the same rate.

A second way to enhance interpretability of quasi-experimental designs is to create research hypotheses that predict complex interaction effects. If the data fall into expected patterns, it often becomes difficult to develop a rival alternative hypothesis that adequately describes the data, if the complex interaction is anticipated. This approach provides strong suggestive evidence that the proffered hypothesis is supported, and the causal dynamics that formed the basis of the (quasi-experimental) design deserve close, open-minded attention.

Case-Control Designs

Case-control designs are based on data already in hand, which is why they sometimes are called *ex post facto* or retrospective designs. The goal of such designs is to address an issue that almost always precludes experimentation, for reasons of ethics or feasibility. The study method involves identifying hypothesized causes of an outcome retrospectively, which usually were not even considered when the suspected cause occurred. An example of this type of design is found in a study of unexpected cancers arising in women between 15 and 22 years of age between 1966 and 1969 in the same hospital in Boston, MA. The young women were diagnosed with adenocarcinoma of the vagina, which usually is seen in women over the age of 50 (See Goodman et al., 2011). This aberrant trend stimulated medical researchers to seek a reason. After considerable effort, a likely candidate emerged – diethylstilbestrol or DES – a synthetic form of estrogen that had been prescribed to pregnant women between 1940 and 1971 to prevent miscarriage and related complications of pregnancy.

Obviously, the hypothesis that DES was to blame could not be tested experimentally. However, considerable data were available to prompt researchers to investigate this hypothesis retrospectively. In an ingenious study, Herbst and colleagues (1971) identified eight women who had contracted vaginal cancer in their teens and early 20s, 19 of whom were delivered in the same hospital (Vincent Memorial Hospital in Boston, MA) and the last at a nearby hospital, though at different times (all were born between 1946 and 1951). The investigators paired each of the women with four other women who had been born in the same hospital within a week of the birth of the critical subjects. They compared the rates of cancer between the two groups. The main results of the case-control study are shown in Table 9.2.

As can be seen immediately, the odds of contracting vaginal cancer in the women whose mothers were *not* treated with DES were extremely low, whereas for those whose mothers were treated with the drug, the odds are frighteningly high. The difference between the two groups was statistically significant at $p < .00001$. Even so, care must be taken in drawing the conclusion that DES *caused* this catastrophic outcome. Other factors might have made a difference, and so, other possible causal factors that might have differentiated the groups were studied: They included differences in birth weights, complications of delivery, ingestion of other drugs during pregnancy, childhood tonsillectomies, household pets, smoking, parents' alcohol use, and SES of parents. Any of these variables might have differentiated the two groups – but they did not. Herbst et al. (1971) considered all these factors, and many others, and found only two other differentiators, bleeding during pregnancy and prior pregnancy loss. These two factors, which might have contributed to the differences that were discovered, were related to the outcome, but only weakly in comparison with the significant effect of the DES-treated versus no DES control group. Years later, the National Cancer Institute's "Third Generation Study" of the afflicted women showed the continuing health problems linked to DES (Titus et al., 2019).

Imagine you were the head of the FDA, and in a position to demand the drug be withdrawn when the Herbst et al. paper was published in 1971. Recall, however, you have no experimental

TABLE 9.2 Results of a case-control study of adenocarcinoma of young women whose mothers were treated with DES in pregnancy

<i>Does daughter have the indicated cancer?</i>	<i>Did mother take DES?</i>	
	<i>Yes</i>	<i>No</i>
Yes	7	0
No	1	32

data, the gold standard of causal inference, to rely on. And yet, the results seem so strong. Ultimately, the FDA recommended doctors desist DEA use, but did not ban the drug, which continued to be used, thought at considerably lower rates, into the 1980s.²

Time-Series Designs

Problems associated with sources of nonequivalence make many non-random comparison group designs inadequate with respect to internal validity. Accordingly, methodologists have searched for other comparison methods that replace or supplement comparison groups as a basis for assessing change. The need for alternatives to comparison group designs is especially acute when social interventions that affect an entire population (a school, city, state, nation) are introduced all at once, and their effects studied over time. In this case, a popular method for assessing change involves a comparison of aggregated observations before and after the treatment is introduced. Consider, for example, a measure of the incidence of violent crimes in a state for the year before and the year after the introduction of a moratorium on capital punishment. A single pretest-posttest assessment of change is impossible to interpret without some knowledge of the degree of fluctuation expected between the measures in the absence of any true change.

The hypothetical crime statistics represented in Figure 9.7 illustrate the problem. Suppose Tennessee had outlawed capital punishment in 2013. The effects of this change in the next year might be as shown in Figure 9.7. The rate of crimes that carried the punishment of death found in the year after the law's change may be interpreted in several different ways. The data may represent an actual increase in these crimes when capital punishment was removed as a deterrent. On the other hand, it may simply reflect the normal year-to-year fluctuation in crime rates which, by chance, have increased over the period studied.

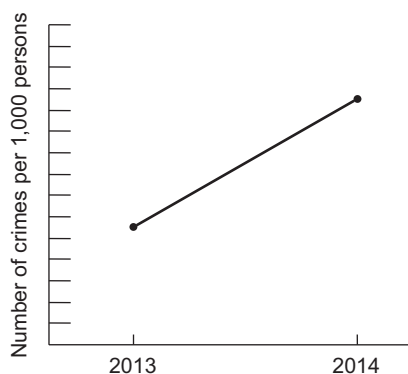


FIGURE 9.7 Hypothetical data illustrating a change of capital crime rates following the abolition of capital punishment in the earlier year. Note that the interpretation of the year-to-year change is almost always ambiguous with trend analyses involving only two data points.

To make matters worse, social quasi-experiments such as this one often are introduced under exceptional social conditions. Ameliorative efforts are more likely to be undertaken when undesirable events have reached a peak or, in the case of our present example, public opinion may have been particularly amenable to an experiment in eliminating capital punishment following a period of exceptionally low homicides. If this is the case, differences in measures taken before and after the quasi-experimental intervention may simply reflect regression back to normal rates, or a continuance of an ongoing secular trend as in Figure 9.8.

In an **interrupted time-series design**, the relative degree of change that occurs after a quasi-experimental treatment may be examined by comparing observations taken at multiple time points prior to the intervention with observations of many time points occurring after intervention. Interpreting the two-round data from Figure 9.7 could be different depending on what trends preceded the intervention. Time-series data might suggest that the 2013–2014 change in capital crime violations might represent normal year-to-year fluctuation in the crime index, with no appreciable rise above previous years. Other data might indicate the results were consistent with a general trend toward year-to-year increases established long before the introduction of the treatment; or the year 2013 (pre-intervention) might have been a particularly low year, with 2014 representing a slight move back toward previous rates. Or, as in Figure 9.9, the

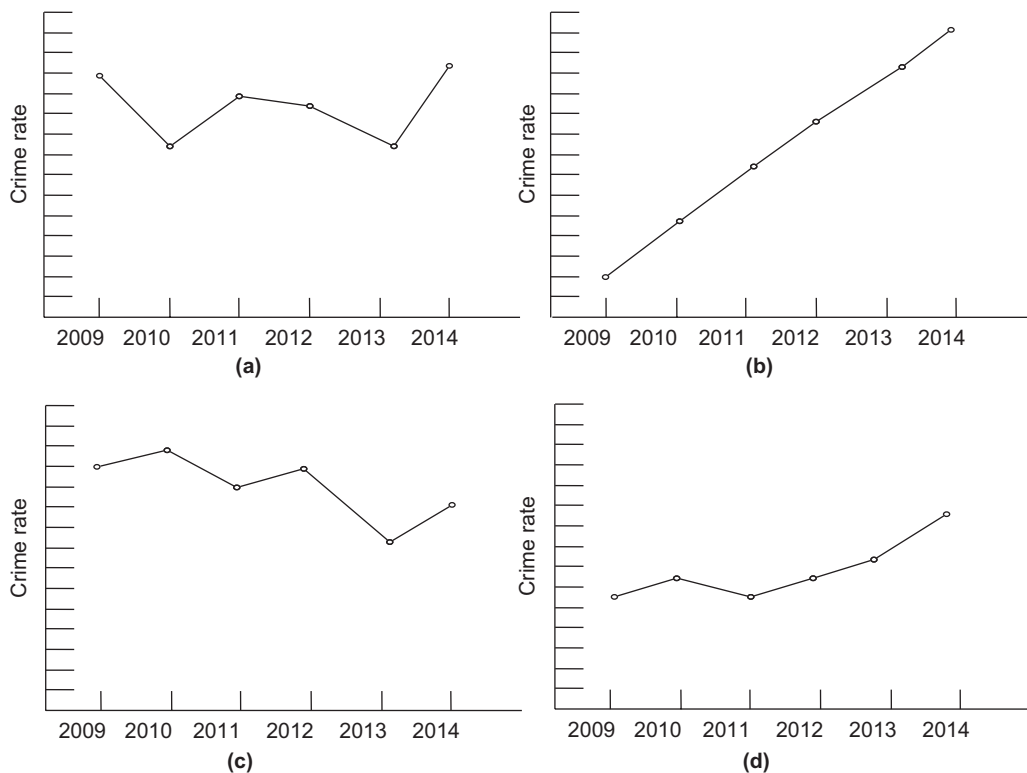


FIGURE 9.8 Possible time trends by year. Note how differences in prior years' rates may affect interpretation of data following the change of a relevant law or rule and may alter evaluations of the change's effects. In Figures 9a and 9c, the rate varied dramatically, but in Figures 9b and 9d, the rates seemed a continuance of existing trends.

analysis might signal a change in direction from the normal pattern, indicating that a meaningful change had occurred in the rate could not be made with much confidence. In all these cases there is reason to believe that the 2013–2014 change did not occur.

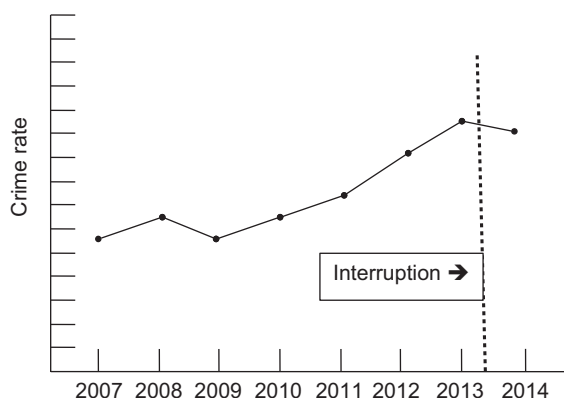


FIGURE 9.9 Interrupted time-series design, with interruption between years 2013 and 2014. The downward trend occurring after the interruption, in contrast to the continual growth over the prior years, might give rise to hope in the effect of the interruption, but lacking more observations over the years post-intervention, such inferences should be tempered.

We would be more certain of the meaningfulness of the change if the time series were measured for more years after the intervention to determine whether the change was stable. Such an interrupted time-series design with extended time points is more representative of time-series results reported in the scientific literature. However, knowing that a meaningful and stable change in scores has occurred in a time-series analysis does not rule out sources other than the intervention as possible causes of the change. Alternative explanations might be available, such as an abrupt increase in population density, changes in recordkeeping procedures, an economic downturn, or other factors related to crime rate that could have occurred simultaneously with the time at which the treatment was implemented.

Statistical analyses of changes in time-series data are afflicted by two major problems. One is that “errors” (i.e., extraneous unmeasured factors) that influence the data obtained at any specific time point tend to be correlated with measurements collected at adjacent time points. These **autocorrelated errors** occur because random events that affect the measurement obtained at one time is more likely to carry over and be correlated with measurements taken at temporally adjacent points than with observations separated by longer time delays between assessments. Such carry-over errors make it more difficult to pinpoint a change in the time series at the one specific time of interest to the evaluation researcher. For example, criminal activity and crime patterns in year 2014 are more likely to be autocorrelated with those in 2013 than with those of 2024.

The second problem that affects time-series analyses is the presence of *systematic trends* or *cycles* that affect the pattern of data over a specific time and are unrelated to the intervention. Changes due to the treatment of interest must be separated from normal changes that occur cyclically across time. When data are obtained monthly, for instance, regular seasonal fluctuations that may operate across the year must be considered. Some social indicators, for example, appear influenced by weather conditions. Such patterns introduce complications in the analyses

of time-series designs, but they are not impossible to correct. Statistical procedures known as “prewhitening” can be applied to remove regularities in the time series before analyses of experimental effect are begun (Box & Jenkins, 1970).

An applied example of research that made use of an interrupted time series, and whose data were affected by both autocorrelated error and systematic trends is provided in Ramirez and Crano’s (2003) study of California’s “three-strikes” law, which took effect in California in 1994. This law made a 25 year-to-life sentence mandatory for anyone convicted of a third major felony. Ramirez and Crano questioned whether criminals really calculated the cost/benefit ratio before committing a crime. If they did, the three-strikes law would serve as a deterrent. If not, then the law would not preclude bad behavior, but would systematically lower the long-term prospects of such crimes occurring, because the most active criminals account for a large proportion of all crimes that are committed (Blumstein, 2016; Sullivan & Piquero, 2016). To test this possibility, the researchers used archived data to analyze monthly crime rates occurring in the 12 years before and five years after the implementation of the law. Crimes are cyclical in nature – for example, in each year of the study, crime spiked in December. (We will not speculate on the reasons for this, but the result certainly gives little comfort to those singing of peace on earth, good will toward men.) In addition, the data were obviously autocorrelated. Numbers of crimes committed in June were more closely correlated to those of May and July than those of February and April.

The statistical procedures suggested by Box and Jenkins (1970) and Berry and Lewis-Beck (1986) supplied the necessary corrections for cyclicity and autocorrelation effects. The time-series analyses provided a plausible answer to the question, “Does the law work?” but the answer was nuanced, depending on the type of crime involved. When studying crimes of passion (e.g., violent physical assaults), the three-strikes law reduced the violent crime rate over the long run, but it had no immediate impact. This result indicates the law operated not as a deterrent, but rather because it took violent career criminals off the streets (criminologists call this *incapacitation*).³ For nonviolent, “white collar” crimes, however, the three-strikes law appeared to have both an immediate deterrent and a more long-term incapacitating effect. That is, it was related to an immediate and statistically significant decrease in white collar crime from the measurements taken in the months immediately before and immediately after the intervention, while at the same time incarcerating those who did engage in these activities and were caught. Both effects cumulated, resulting in a dramatic decline in white collar crime. However, the analysis disclosed no relation whatever between the three-strikes law’s deterrence or incapacitation when drug-related crimes were involved. This suggests an apparently inexhaustible supply of potential drug kingpins willing to step in and take the place of their incapacitated predecessors.

Comparison time-series designs. Assuming comparable statistical records are available across times and places, **comparison time-series designs** combine features of an interrupted time-series design and a comparison group design. If a social program is introduced in one location or institution but not in some other, preexisting differences between the treatment and comparison site make it difficult to interpret any post-treatment differences. However, if time-series data based on the same recordkeeping system are available for both sites, and if both are subject to similar sources of cyclical and noncyclical fluctuations, then the time-series data from the comparison control group can serve as an additional baseline for evaluating differences in change in the experimental series. When the time series of two groups are roughly parallel in trends of scores gathered prior to the introduction of the intervention, but diverge significantly afterward, many potential alternative explanations for the change in the latter series can be ruled out. As with any time-series design, statistical analyses of a comparison time-series design

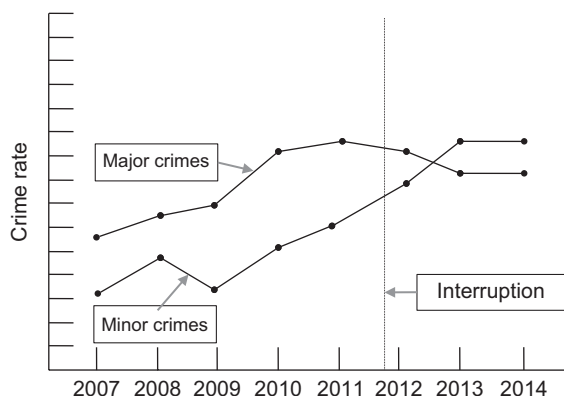


FIGURE 9.10 Comparison time-series data. The figure shows the apparent effect of a three-strikes law on major crimes, but no effect on a comparison data series concerned with crimes not covered by the three-strikes law.

can be complex (Berk et al., 1979; Berry & Lewis-Beck, 1986), but the logic of the design is straightforward, and has been put to good use in many evaluations.

A second method of forming comparisons for the interrupted time series is to include variables in the analysis that are like those in the focal analysis but should not be affected by the interruption. Ideally, the interrupted data (i.e., major “striking” crimes) should be affected, whereas those not subject to the interruption (i.e., petty crimes) should not. These crimes should have been affected by the same extraneous factors as serious crimes (e.g., general economic conditions, social unrest, etc.), but not by the three-strikes law, which was directed to only the most serious crimes, not misdemeanors. Thus, ideally, the law’s passage would affect the serious crime rate, but not the petty crime rate, as shown in Figure 9.10. If these differences in these two crime rates occurred after the introduction of the three-strikes law, they would lend support to the efficacy of the law change on serious criminal behavior.

Regression Discontinuity Designs

The final quasi-experimental design to be described here is applicable to cases in which exposure to a treatment program is based on some clear selection principle. Unlike time-series designs, which may involve observations across many time points for one or many comparison groups, the regression discontinuity design involves comparing groups on either side of a cutoff point. **Regression discontinuity (RD) designs** are conducted to test the existence of some systematic relationship between a pretest selection variable, which is used to place participants in comparison groups, and a posttest measure. If individuals are selected for inclusion in a special educational enrichment program based on achievement test scores, for instance, we would expect their test scores (apart from any program effects) to be positively linearly related to later measures of educational attainment. The RD design is used to determine if there is a discontinuity in score distributions between those immediately above and immediately below the cutoff for inclusion in the program, and differences in the slopes of the regression lines on each side.

The RD design is quasi-experimental because it is meant to mimic a true experiment in which a group of participants at a cutoff point are randomly assigned to a treatment or a control condition. For example, suppose the State Department of Education has developed a program for children whose families fall below the poverty line established by the federal government.

To test the effect of the program, the state might take those falling slightly above and slightly below the poverty line and *randomly assign* them to the program or control condition. The effect of the program then could be tested in a similar way to the pretest-posttest control group experimental design. Such a design might prove politically unfeasible, however, and the RD design provides a means of testing program effects without making assignments that are difficult to justify on any basis other than scientific utility.

In the RD approach, if selection into the special program is associated with a specific cut-off point on the selection variable, we can use all of those who fall below the cutoff point as a comparison to those falling above the cutoff. Although overall we would expect those below to perform differently from those above the cutoff, in the absence of a special treatment, we would not expect any abrupt change in outcomes for those falling near one another on either side of the cutoff. If such a change did occur, we could take this as evidence of a treatment effect above and beyond the expected selection effect (Thistlethwaite & Campbell, 1960). This design is especially useful when scarce resources are used as interventions. Such a *regression discontinuity* effect is illustrated in Figure 9.11 for the case of a positive (merit-based) selection factor. A full treatment of the assumptions and statistical analyses for this design is available in Reichardt and Henry (2012) and Hallberg et al. (2013).

The RD design is less common than comparison time-series designs in program evaluation settings, partly because assignment of participants to groups based on a criterion cutoff score is relatively rare, and partly because it requires extensive data collection for individuals across the full range of scores on the selection variable. However, when RD is used, the analysis often is focused on important, socially relevant issues. For example, using this approach, researchers have investigated the effects of mandatory divorce counseling laws (Deluse, 1999), HIV prevention programs in West Africa (Arcand & Wouabe, 2010), targeted funding for disadvantaged

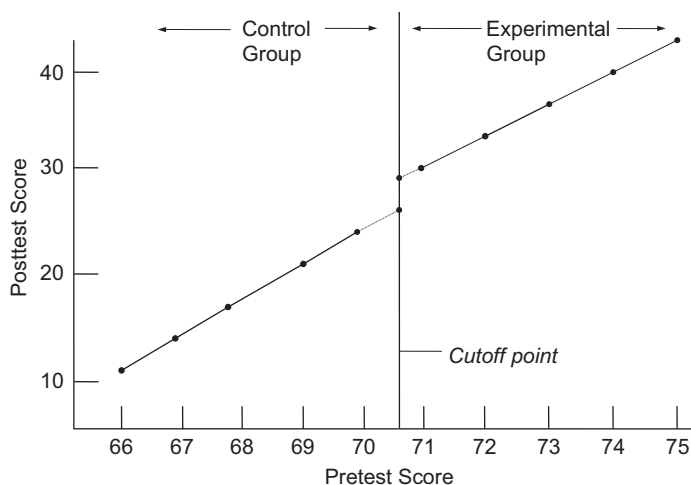


FIGURE 9.11 Regression discontinuity design. Those above the cutoff point are designated for special treatment (e.g., financial help for the most highly gifted to attend college, as is done with National Merit Scholarships, or special advanced programs for gifted and talented students), or special help for those in exceptional need. The interruption is viewed as an intervention, targeted specifically on a designated group. The analysis is meant to test the effect of the intervention on the posttest score, indicated by a significant discontinuity in the regression line, as shown here.

children (Henry et al., 2010), and programs for mathematically gifted black children (Robinson et al., 1990). Other findings of highly relevant issues for youth involve effects of lowering the age of legal handgun purchase from 21 to 18 years. The result of a RD analysis by Raifman and colleagues (2020) showed that a contemplated change in state policies that would allow 18-year-olds to buy handguns (formerly, the legal age was 21 years of age) had resulted in an increase of more than 300 suicides by adolescents in states that had adopted an 18-year age limit to buy handguns. In another important domain, Bell (2021) found via RD analysis that free community college enhanced students' success in admission to four-year colleges and resulted in higher GPAs than those who had not attended the available free community colleges. This result indicated the program developed in Tulsa, OK, was indeed worth the investment, a highly contested issue in that community.

RD is an interesting quasi-experimental design to consider, especially in contrast to true experimental designs. Whereas true experiments are based on randomization as the basis for assignment to different groups, the regression discontinuity design is based on a *non-random* cutoff score. However, only strict application of the RD selection rule permits cause-effect conclusions to be drawn. In addition, it is essential that the relationship between pretest and posttest is linear, or the interpretability of the ensuing results is compromised. A downside of RD designs is that the statistical power of such analyses is low: Using RD analysis requires two to four times as many observations as experimental designs to attain the same degree of statistical power. Given these problems, RD studies obviously are not preferable to randomized experiments, if the latter is possible. However, in special cases, RD research designs help illustrate how research purposes can be adapted to policy-relevant conditions without significant loss of causal interpretability. In the appropriate circumstances, RD designs provide information that may prove extremely useful for policy-making purposes.

The Use of Archival Data in Longitudinal Research

Researchers rarely have sufficient lead-time prior to the introduction of an intervention to obtain premeasures over an extended period specifically for the purposes of evaluation. Hence, using time-series quasi-experimental designs often depends on the availability of statistical records or other archival documents. Sometimes the only relevant and available historical materials are written records (newspaper summaries, case records, personal letters, etc.) that must be subject to *content analysis* to be useful as research data. (The methods and purposes of content analyses are covered in Chapter 14.) For present purposes we have dealt only with research based on the availability of quantitative records or statistical indices compiled by institutions or government agencies for various accounting purposes.

Fortunately, North America and Europe have proven “statistics happy” for much of their recent histories. In the United States, for example, in addition to the constitutionally mandated census of the entire population every ten years, which provides valuable demographic information on a regular basis, numerous federal agencies are charged with maintaining statistical databases of records of births and deaths; hospital admissions and other health records; various indices of economic activity; records of arrests, convictions, and other indices of criminal activity; unemployment statistics; and the like. Parallel recordkeeping goes on at state and local levels, which is important for evaluations of locally implemented social programs.

Use of statistical archives has many advantages, but also creates some disadvantages for research purposes. First, it limits the dependent measures or outcome variables collected, and this can limit the kinds of questions that can be asked. Records kept for purely administrative purposes often do not reflect the primary goals of the social program being evaluated – Did the

program work as opposed to the number of people served? In evaluations of the criminal justice system, it is easier to obtain archival records on recidivism (re-arrests or re-imprisonment) than on a program's positive outcomes. In educational settings, information on achievement test results is more likely to be available than are indicators of other types of positive student outcomes such as social adjustment or moral or emotional development.

Another limitation imposed by reliance on archival records is the time interval covered by a given archive, which may not be the unit of time ideal for research purposes. If statistics are accumulated daily, researchers have considerable flexibility to aggregate information over any period they choose, though they must contend with the costs involved in compiling large data-sets. On the other hand, if summary statistics have been maintained only on a yearly basis, the research will have to cover a long period for enough data points to be available for statistical analyses.

Finally, a major worry with many archival studies is the possibility that the nature or method of recordkeeping has changed over time. Recordkeeping systems can be altered in many ways, often for administrative convenience. For one thing, the *criterion* for inclusion in the data file may be changed. Crime statistics, for instance, can be dramatically affected by changes in the activities or actions on which police are required to file reports. Records also may be altered in form or content, such as changes in categories of information or in times at which information is recorded. Sometimes, and this is especially relevant in medical research, new techniques of diagnosis will allow more certain or more rapid identification of diseases that could not be diagnosed clearly before the new test. This change may produce data that suggest a rapid rise in the incidence of a disease. Sometimes the researcher will know enough about the diagnosis or recordkeeping system to make conversions between different versions of the same information, but more often such changes render data noncomparable from one time to another. If changes in administrative recordkeeping methods occur near the introduction of a social program, the records may become useless as measures of program-related changes.

On a positive note, use of archived data usually provides much broader swaths of the population than is possible in laboratory-oriented research. "Big data" is here to stay. Often, archival data are representative of large groups of the populace – cities, states, the country. Being able to extrapolate one's results to the entire population of, say, adolescents across the country, or pensioners in Ohio, or all social welfare recipients in West Virginia lends value to the research enterprise. It helps offset the difficulties and restrictions that the quasi-experimental methods can impose on the researcher and, for some, makes the trade-off worthwhile.

Conclusion

Quasi-experimental methods serve as a bridge between nonexperimental and experimental methods. The quasi-experimental approach allows us to study important issues and to make estimates of the strength of manipulated or naturally occurring "treatments." The confidence we can place in these estimates usually is not as great as that which we can derive from true experimental designs, which involve random assignment, but sometimes the issue is important enough that we are willing to pay this price to generate estimates of the strength of relations, or the effectiveness of treatments. With sufficient thought and effort, quasi-experimental designs can produce important insights which, because of contextual constraints, might not have been researchable through the more standard experimental techniques. In this chapter, we have only scratched the surface of the variety of possible quasi-experimental designs and their application. The chapter was meant to lay the groundwork. The elaboration of the many variations that can be implemented is a function of the reader's ingenuity, creativity, and motivation.

Notes

- 1 A critic might argue that children in early morning classes are more energetic learners than later in the afternoon, and this difference accounted for the difference in learning ability in the subsequent measurement of year 4.
- 2 For a depressing timeline of the DES trajectory see DES Action's *DES TIMELINE*, at <https://desaction.org/des-timeline/>.
- 3 Criminologists suggest that 90% of all crimes are committed by 10% of the criminal population. As such, Incarcerating a "career criminal" (incapacitation) will have a disproportionately large effect on crime rates, and the California law did just that.

References

- Arcand, J.-L., & Wouabe, E. D. (2010). Teacher training and HIV/AIDS prevention in West Africa: Regression discontinuity design evidence from the Cameroon. *Health Economics*, 19(Suppl 1), 36–54.
- Arceneaux, K. (2005). Using cluster randomized field experiments to study voting behavior. *Annals of the American Academy of Political and Social Science*, 601, 169–179.
- Bell, E. (2021). Does free community college improve student outcomes? Evidence from a regression discontinuity design. *Educational Evaluation and Policy Analysis*, 43(2), 329–350.
- Berk, R. A., Hoffman, D. M., Maki, J. E., Rauma, D., & Wong, H. (1979). Estimation procedures for pooled cross-sectional and time series data. *Evaluation Quarterly*, 3, 385–410.
- Berry, W. D., & Lewis-Beck, M. S. (1986). *New tools for social scientists: Advances and applications in research methods*. Beverly Hills, CA: Sage.
- Blumstein, A. (2016). From incapacitation to criminal careers. *Journal of Research in Crime and Delinquency*, 53(3), 291–305.
- Boruch, R. (2012). Deploying randomized field experiments in the service of evidence-based crime policy. *Journal of Experimental Criminology*, 8(3), 331–341.
- Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis: Forecasting and control*. San Francisco, CA: Holden-Day.
- Campbell, D. T., & Erlebacher, A. (1970). How regression in quasi-experimental evaluation can mistakenly make compensatory education look harmful. In J. Hellmuth (Ed.), *The disadvantaged child (vol. 3). Compensatory education: A national debate* (pp. 185–210). New York: Brunner-Mazel.
- Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp 171–246). Boston, MA: Houghton Mifflin.
- Crano, W. D., Alvaro, E. M., & Siegel, J. T. (2015). The media campaign as a focal prevention strategy: A guide to their design, implementation, and evaluation. In L. M. Scheier (Ed.), *Handbook of drug prevention* (pp. 397–414). Washington, DC: American Psychological Association.
- Crano, W. D., & Messé, L. A. (1985). Assessing and redressing comprehension artifacts in social intervention research. *Evaluation Review*, 9, 144–172.
- De Dominicis, S., Crano, W. D., Ganucci Cancellieri, U., Mosco, B., Bonnes, M., Hohman, Z., & Bonaiuto, M. (2014). Vested interest and environmental risk communication: Improving willingness to cope with impending disasters. *Journal of Applied Social Psychology*, 44(5), 364–374.
- De Dominicis, S., Ganucci Cancellieri, U., Crano, W. D., Stancu, A., & Bonaiuto, M. (2021). Experiencing, caring, coping: Vested interest mediates the effect of past experience on coping behaviors in environmental risk contexts. *Journal of Applied Social Psychology*, 51(3), 286–304.
- Deluse, S. R. (1999). Mandatory divorce education: A program evaluation using a 'quasi-random' regression discontinuity design. *Dissertation Abstracts International: Section B: the Sciences & Engineering*, 60, 1349.
- Galton, F. (1885). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute*, 15, 246–263.
- Goodman, A., Schorge, J., & Greene, M. F. (2011). The long-term effects of in utero exposures – The DES story. *New England Journal of Medicine*, 364, 2083–2084.

- Green, D. P., Aronow, P. M., Bergan, D. E., Greene, P., Paris, C., & Weinberger, B. I. (2011). Does knowledge of constitutional principles increase support for civil liberties? Results from a randomized field experiment. *The Journal of Politics*, 73, 463–476.
- Hallberg, K., Wing, C., Wong, V., & Cook, T. D. (2013). Experimental design for causal inference: Clinical trials and regression discontinuity designs. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods: Foundations*. (Vol. 1, pp. 223–236). New York: Oxford University Press.
- Henry, G. T., Fortner, C. K., & Thompson, C. L. (2010). Targeted funding for educationally disadvantaged students: A regression discontinuity estimate of the impact on high school student achievement. *Educational Evaluation and Policy Analysis*, 32, 183–204.
- Herbst, A. L., Ulfelder, H., & Poskanzer, D. C. (1971). Adenocarcinoma of the vagina. Association of maternal stilbestrol therapy with tumor appearance in young women. *New England Journal of Medicine*, 284(15), 878–881.
- Hovland, C. I., Lumsdaine, A. A., & Sheffield, F. D. (1949). *Experiments on mass communication*. Princeton: Princeton University Press.
- Jilke, S., Lu, J., Xu, C., & Shinohara, S. (2019). Using large scale social media experiments in public administration: assessing charitable consequences of government funding of nonprofits. *Journal of Public Administration Research and Theory* 29(4), 627–639.
- Kershaw, D., & Fair, J. (1976). *The New Jersey income-maintenance experiment*. New York: Academic Press.
- Krueger, A. B., & Zhu, P. (2004). Another look at the New York city school voucher experiment. *American Behavioral Scientist*, 47, 658–698.
- Levine, R. A., Watts, H., Hollister, R. G., Williams, W., O'Connor, A., & Wilderquist, K. (2005). A retrospective on the negative income tax experiments: Looking back at the most innovative field studies in social policy. In K. Wilderquist, M. A. Lewis, & S. Pressman (Eds.), *The ethics and economics of the basic income guarantee* (pp. 95–106). Farnham: Ashgate.
- McNemar, Q. (1940). A critical examination of the University of Iowa studies of environmental influences upon the I.Q. *Journal of Psychology Bulletin*, 37, 63–92.
- O'Leary, A., Ambrose, T. K., Raffaelli, M., Maibach, E., Jemmott, L. S., Jemmott, J. B. III, Labouvie, E., & Celentano, D. (1998). Effects of an HIV risk reduction project on sexual risk behavior of low-income STD patients. *AIDS Education and Prevention*, 10, 483–492.
- Raifman, J., Larson, E., Barry, C. L., Siegel, M., Ulrich, M., Knopov, A., & Galea, S. (2020). State handgun purchase age minimums in the US and adolescent suicide rates: Regression discontinuity and difference-in-differences analyses. *BMJ: British Medical Journal*, 370, 1–8.
- Ramirez, J. R., & Crano, W. D. (2003). An interrupted time series analysis of California's three strikes law on instrumental, violent, minor, and drug-related crime: Deterrence and incapacitation. *Journal of Applied Social Psychology*, 33, 110–144.
- Reichardt, C. S., & Henry, G. T. (2012). Regression-discontinuity designs. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological* (pp. 511–526). Washington, DC: American Psychological Association.
- Robinson, A., Bradley, R. H., & Stanley, T. D. (1990). Opportunity to achieve: Identifying mathematically gifted black students. *Contemporary Educational Psychology*, 15, 1–12.
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.
- Salovey, P., & Williams-Piehot, P. (2004). Field experiments in social psychology: Message framing and the promotion of health protective behaviors. *American Behavioral Scientist*, 47, 488–505.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton: Mifflin and Company.
- Sullivan, C. J., & Piquero, A. R. (2016). The criminal career concept: Past, present, and future. *Journal of Research in Crime and Delinquency*, 53(3), 420–442.
- Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51, 309–317.
- Thorndike, R. L. (1942). Regression fallacies in the matched groups experiment. *Psychometrika*, 7, 85–102.
- Titus L., Hatch, E., El, Drake, K. M., Parker, S. E., Hyer, M., Palmer, J. R., Strohsnitter, W. C., Adam, E., Herbst, A. L., Huo, D., Hoover, R. N., & Troisi, R. (2019). Reproductive and hormone-related

- outcomes in women whose mothers were exposed in utero to diethylstilbestrol (DES): A report from the US national cancer institute des third generation study. *Reproductive Toxicology*, 84, 32–38.
- Vonderlin, R., Müller, G., Schmidt, B., Biermann, M., Kleindienst, N., Bohus, M., & Lyssenko, L. (2021). Effectiveness of a mindfulness- and skill-based health-promoting leadership intervention on supervisor and employee levels: A quasi-experimental multisite field study. *Journal of Occupational Health Psychology*, 26(6), 613–628.

10

NONEXPERIMENTAL RESEARCH

Correlational Design

Previous chapters of this book were focused on considerations relevant to the development and application of experimental designs. In these designs, participants' exposure to a treatment or manipulation is controlled by the researcher. The major purpose of such control is to reduce or eliminate the plausibility of confounds or alternative explanations for effects on the dependent variable. In experimental research, a clear distinction is made between the independent (or causal) variable, which is controlled or manipulated by the experimenter, and the dependent (outcome) variable, which is allowed to vary freely. In many areas of research, however, experimental control over variables is impossible, unethical, or impractical.

A researcher studying the effects of demographic characteristics such as sex, age, height, or religious affiliation, for example, is in no position to manipulate these variables. People possess predetermined levels of these characteristics, and these features of the individual are beyond the immediate influence of the researcher. Similarly, if respondents' levels of a variable are determined by their responses on a measuring instrument, such as a scale of authoritarianism, extraversion, or need for cognition, the researcher again is not able to manipulate this variable. These are intrinsic and generally unmalleable features of the people. In most investigations of this type, the research question becomes focused on the extent that variables are in some way related to or associated with one another. The type of analysis thus becomes nonexperimental or correlational. By correlational methods, we are referring to a variety of techniques that include the simple correlation between two variables, regression with a single predictor, multiple regression, and other approaches used to examine statistical relations among variables in non-experimental research.

In a correlational or nonexperimental research design, the variables are allowed to vary freely, and the researcher observes and records the extent of their covariation, that is, the extent that variations in one variable are associated with (but not necessarily caused by) variations in another variable, or vice versa. The most basic correlational study is one in which the association between two variables is investigated. Higher scores on a questionnaire to measure religiosity, for instance, may be associated with higher scores on an instrument designed to measure authoritarianism. Participants' responses on both measures are not manipulated or controlled by the researcher – the question is the extent that scores on one measure vary together with scores

on the other. Does a particular individual's score on one measure give us a clue to, or help us predict, that individual's response to the second measure?

A major advantage of nonexperimental or correlational research is that it allows both variables of interest to vary freely, so that the direction and degree of relation can be determined. In most correlational research, a variable may be measured along some continuum involving quantitative response options. However, not all nonexperimental research takes advantage of the full continuum of variations in the measured variables. Consider a study in which participants with varying degrees of antisemitism were assessed. Like most attitudinal variables, antisemitic prejudice can be measured on a quantitative scale, and this allows the possibility of identifying variation across participants on this measure. However, the researcher may choose to use this measure to divide participants into discrete groups or categories for the purpose of comparison. This necessitates redefining variation among individuals according to some categorical scheme, such as "high" or "low." Thus, for purposes of the research design, individuals with similar scores on antisemitic sentiment will be considered members of the same group.

Such categorization can result in an unnecessarily imprecise use of information. For example, consider the data in Table 10.1. This illustration presents the scores of ten participants who were administered a scale of antisemitism, arranged in order from lowest to highest scores. Suppose the researcher wished to categorize these participants into two groups of equal size. One possible decision rule is to use a **median split**, a process in which participants are divided at the "middlemost" score, with 50% of participants above the median termed the "high antisemitism group," and the 50% below the median the "low antisemitism group."

As can be seen in Table 10.1, dividing participants into discrete classifications based on a quantitative measure has some undesirable features. The four participants nearest the median have very similar scores, yet two of them (Participants 10 and 4) have now been defined (by the median split) as being in the low antisemitism group, whereas the other two (Participants 2 and 5) have been placed in the high group. This classification is applied despite the fact that the scores of Participants 10 and 4 are more similar to those of Participants 2 and 5 (who have been classified in a different group) than they are to the other participants in their own group. Participants with scores of 3 and 50 are classified as being identical in terms of their antisemitism grouping – they are both in the "low" group – whereas those with scores of 50 and 51 are defined as different! When a quantitative variable representing a wide range of scores is transformed to be categorical, the measure is no longer sensitive to detect the entire range of possible scores.

TABLE 10.1 Conversion of quantitative antisemitism scores into high and low antisemitism groups

<i>Participant</i>	<i>Antisemitism score</i>	<i>Dichotomized antisemitism groups</i>
6	3	Low antisemitism
9	28	
1	30	
10	49	
4	50	
<i>Median split</i>		
2	51	High antisemitism
5	53	
8	67	
3	75	
7	88	

After pondering the issues noted in the previous paragraph, if the quantitative variable is to be converted to a categorical variable for the purpose of analysis, the approach should be based on meaningful or theoretical grounds. A measure that asks respondents to report on the number of lifetime sexual partners may later wish to meaningfully classify participants into zero sexual partners versus one or more sexual partners. This route may be pursued if the research objective is to compare virgins and those who have engaged in sex on well-being. In another example, clinical and health measures usually provide a theoretical rationale, based on previous research evidence, regarding the number of symptoms or ailments to be clinically diagnosed with a disorder. For example, questionnaires that assess psychological depression are typically quantitative in assessing the number of symptoms participants are experiencing from a checklist of depressive symptoms (e.g., low mood, diminished interest, fatigue, diminished concentration, slow thoughts, decrease in appetite, feelings of worthlessness), but a minimum of five symptoms might serve as the binary cutoff to be diagnosed with clinical depression or not. Thus, the depression measure may be used as a quantitative variable if the research objective is to examine the number of depressive symptoms or alternatively used as a categorical variable if the objective is to examine whether participants have that diagnosis or not.

Bivariate Correlation and Regression

Several statistical procedures are available to assess the degree of association between two variables, also known as a **bivariate correlation**. The most computed bivariate correlation is the **Pearson correlation** coefficient, which assesses the extent of linear relation between two quantitative variables, that is, the extent that variation in one measure is consistently associated with variation in another measure. A **scatter plot** can be used to graphically represent a correlation between two variables. Figure 10.1 illustrates scatter plots for positive, negative, and zero correlations for two quantitative variables designated X and Y. Each dot on the scatter plot represents a participant's score on the X (horizontal axis) and Y (vertical axis) variables. As such, the total number of dots signifies the number of participants in the sample.

Figure 10.1a illustrates a positive relation for a Pearson correlation coefficient. Individuals who have relatively low scores on the X measure also tend to have relatively low scores on the Y measure, and those with high X scores also tend to have high Y scores. An example of a positive Pearson correlation is the association between the number of hours studying and scores on a research methods exam. Conversely, Figure 10.1b represents a negative Pearson correlation in which relatively high X scores are accompanied by relatively low Y scores, and relatively low X scores are accompanied by relatively high Y scores. An example of a negative correlation is between percent of class absences and scores on a research methods exam. Figure 10.1c shows a zero Pearson correlation in which low or high scores on X are not systematically accompanied by low or high scores on Y. Notice that as the values of X increase (or decrease), the values of Y remain approximately the same. In other words, no systematic association is evidenced between the X and Y variables. An example of a zero or near zero correlation is that involving a person's height in inches and research methods exam scores (assuming that height should be completely unrelated to exam performance).

The Pearson correlation may range from -1.00 to $+1.00$. The correlation coefficient furnishes two pieces of information about an association: direction and strength (effect size). First, the sign of the coefficient indicates the direction of relation (positive, zero, or negative) between X and Y, as described. Second, the closer the correlation coefficient is to $.00$ the weaker the relation, whereas the closer the coefficient is to an absolute value of 1.00 (regardless of sign) the stronger the magnitude of the relation. A coefficient of $.00$ indicates that no relation exists

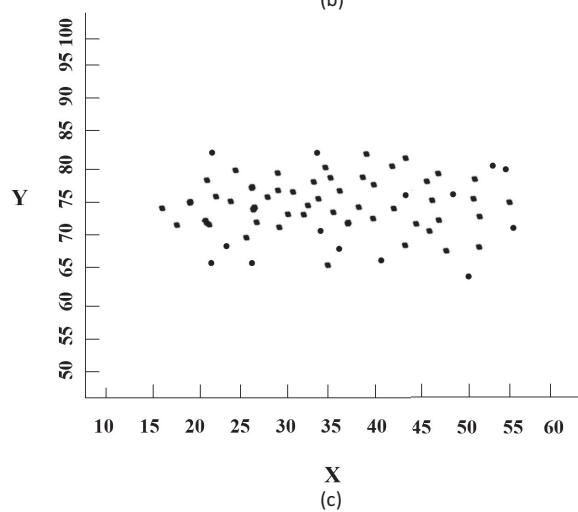
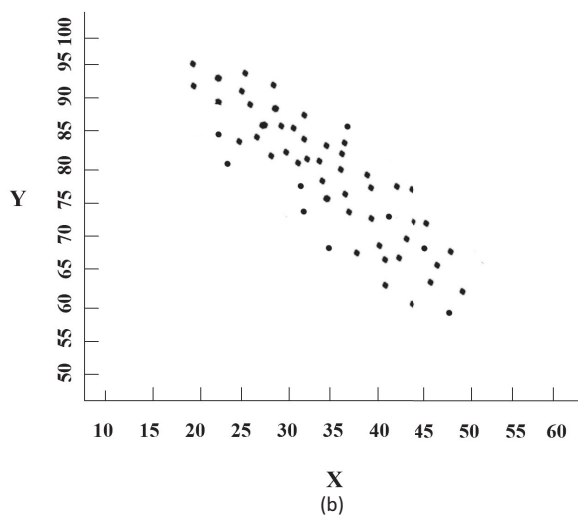
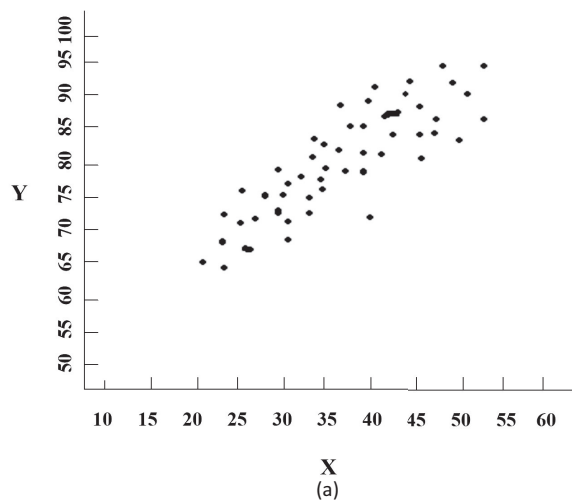


FIGURE 10.1 Scatter plots involving positive, negative, and zero correlations.

between these two variables. A correlation of $+1.00$ represents a perfect positive association and a correlation of -1.00 signifies a perfect negative association. In these instances of perfect correlations, the points in a scatter plot of X and Y scores form a straight line, indicating that the relative standing of any person's score on the X measure corresponds exactly to that person's relative position on the Y measure. An example of a perfect positive correlation ($r = +1.00$) is the association between each person's weight in pounds and weight in ounces. Essentially, these two variables are measuring the same construct of weight but using different measurement metrics that can be directly converted from one to another, so this accounts for their perfect positive correspondence. An example of a perfect negative correlation ($r = -1.00$) involves the association between the percent of class attendances and the percent of class absences for students enrolled in a course. These two variables show perfect negative correspondence as they are assessing the same construct but in opposite ways.

Interpreting a Low Correlation

If the results of a Pearson correlation indicate a low value or value near zero between two variables, four potential explanations for this result are possible. First, and most simply, no systematic relation exists between the two variables.

The second possibility is that a systematic relation exists between the two variables, but the relation is nonlinear. The graph in Figure 10.2 illustrates a nonlinear, specifically a curvilinear relation in which increases in X scores are sometimes accompanied by increases in Y scores and sometimes by decreases. Such a connection may be found between a measure of anxiety and scores on a complex intellectual task such as a research methods exam. At very low levels of anxiety (the X score), due to lack of motivation, fatigue, or boredom, exam performance (Y score) is poor. As anxiety levels (X scores) rise to a moderate level, performance improves until a threshold of high performance is reached. This pattern suggests that a moderate level of anxiety may be motivational in helping with performance. As anxiety increases to a high level, performance decreases due to the debilitating effects of being highly tense and nervous. In this scenario, the linearity assumption for the Pearson correlation is violated, so this statistic should

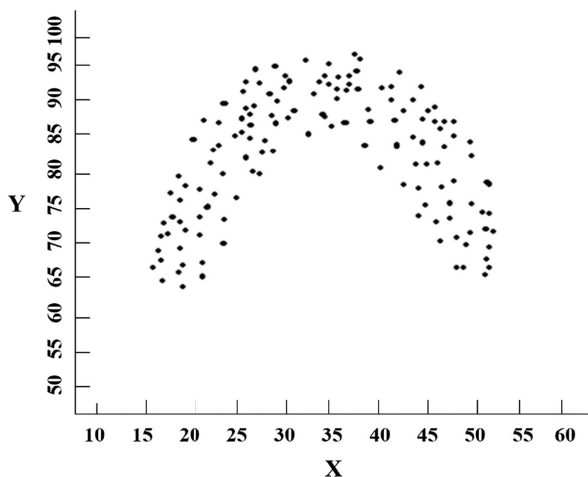


FIGURE 10.2 Scatter plot of a nonlinear relation.

not be used to estimate the association between these two variables. The Pearson correlation coefficient is only appropriate for linear relations involving quantitative variables.

The third possibility is that one or both measures involved in the correlation exhibit poor measurement reliability. Imperfection in measurement always diminishes the apparent relation between variables: the greater the imperfection, the more serious its effect on the correlation coefficient. A highly unreliable measure contains items that do not consistently represent the underlying phenomenon. Thus, such a variable contains only a small reliable proportion of the shared variance, and a large proportion of the measurement is contaminated by measurement error (see the discussion of measurement reliability in Chapter 3). If a Pearson correlation is computed to assess the association between this unreliable measure and another unreliable measure, we are assessing the relation between instruments that contain considerable random noise in the form of measurement error, thereby producing a smaller correlation coefficient.

The fourth possibility is that a very low correlation value may be an artifact of limitations of measurement, known as *restriction of range*. The size of the correlation between any two variables will be automatically attenuated (diminished) if the range of scores on either measure is or both measures are restricted or truncated. A truncated measure, for example, may arise if a self-esteem scale is administered to a sample consisting of people suffering from clinical depression. This is problematic because the possible distributional range of self-esteem scores in these psychologically distressed individuals may be much narrower than that found in the general population. An example of restriction of range is illustrated in Figure 10.3. Figure 10.3a represents the relation between measures on X and Y across a wide range of participant scores on both variables. The Pearson correlation is positive. Figure 10.3b provides a blow-up of the limited range of values represented in the area sectioned off from Figure 10.3a. Within this more restricted range of X scores, the previously positive correlation in the relation between X and Y is no longer evidenced in this subsample. If the correlation coefficient were computed between these two measures for a sample with very little range in their scores on X or Y (Figure 10.3b), the resulting correlation would be much closer to .00 than that of the relation involving the entire distribution of respondent scores (Figure 10.3a). A sufficient range of scores in both variables is necessary to meaningfully compute a Pearson correlation.

Figure 10.4 provides an illustration of how the relation between two variables could be misrepresented if the range of values on one of the variables were unrepresentative of the total range possible. This graph presents, roughly, the association obtained between measures of cognitive complexity (e.g., Streufert et al., 1992) and intelligence. In the lower ranges of intelligence scores, a systematic increase is found in intelligence scores as cognitive complexity increases. However, as the range of intelligence scores reaches a ceiling, as might be expected in a sample of students from a highly selective university, cognitive complexity scores are no longer systematically related to IQ. This attenuation of association using participants at the upper levels of intelligence prompted early investigators (whose participants were primarily college students from highly selective schools) to conclude that cognitive complexity and intelligence were independent. Only after a wider range of intelligence scores was represented in this research did the true nature of the connection become apparent.

Correlation Family: Pearson, Point-Biserial, and Phi Correlations

The discussion of the association between two variables so far has focused on the Pearson correlation, the most widely used variant of the correlation coefficient. Other variants of the bivariate correlation are used when one or both variables are measured with binary response options (i.e., a dichotomous variable). The Pearson correlation is applied when both the X variable and

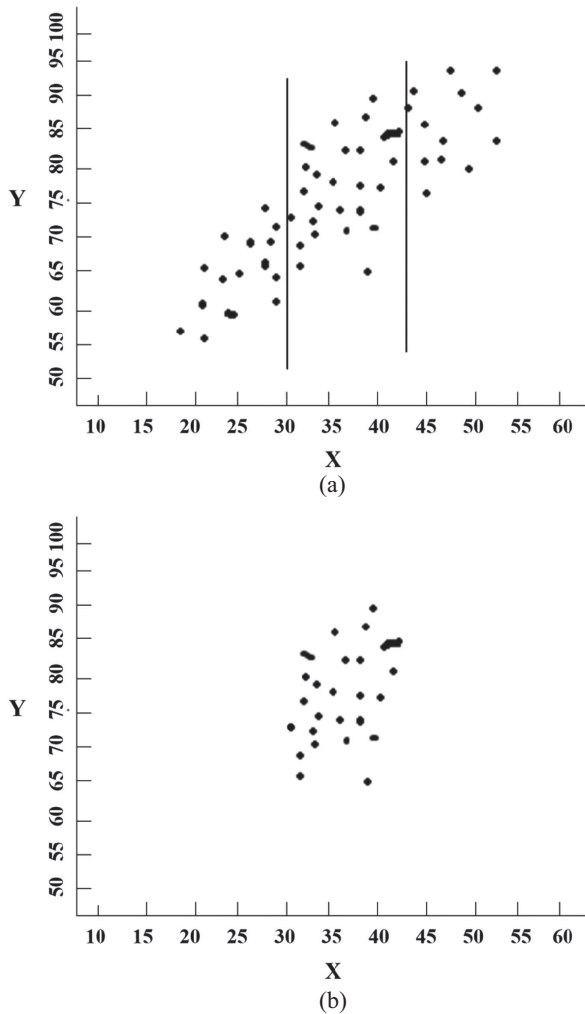


FIGURE 10.3 Example of restriction of range.

Y variable are assessed on quantitative scales. The **point-biserial correlation** (r_{pb}) is applied to examine the association between a binary variable and a quantitative variable. In other words, it is appropriate if the X variable is binary and the Y variable is quantitative or if the X variable is quantitative and the Y variable is binary. The **phi correlation** (ϕ) is applied if the purpose is to examine the association between two binary variables. The Pearson correlation, point-biserial correlation, and phi correlation are all calculated using the same statistical formula. Thus, each could range from a correlation of -1.00 to $+1.00$. The sign of the correlation coefficient offers information about the direction of the relation (positive, negative, or zero) and the absolute value of the coefficient denotes the strength (effect size) of the relation.

Figure 10.5 compares these three types of correlations using different datasets that focus on the same research topic. In the first research scenario (Figure 10.5a), both variables involve quantitative scales. Across participants, those who use marijuana a greater number of times are more likely to believe that marijuana is relaxing. This scenario yields a Pearson correlation of .86.

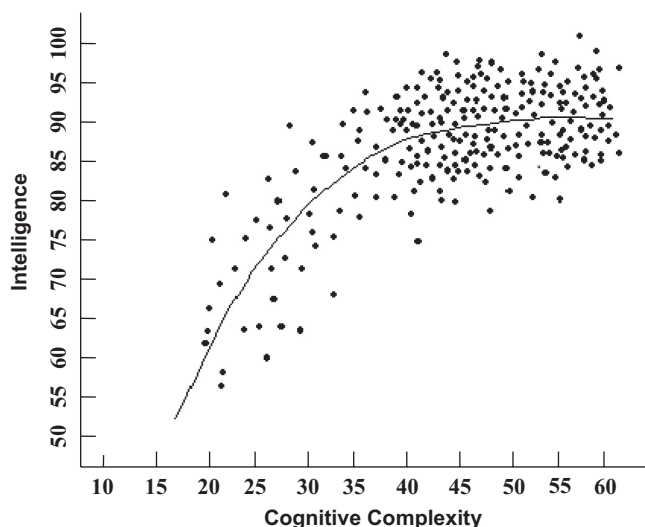


FIGURE 10.4 Nonlinear relation between cognitive complexity and intelligence.

In the second research scenario (Figure 10.5b), one variable is binary and the other variable is quantitative. Notice that as participants shift in their marijuana usage from values of 0 (no) to 1 (yes), they are more likely to endorse higher belief ratings that marijuana is relaxing. This yields a point-biserial correlation of .78. In the third scenario (Figure 10.5c), both variables involve binary response options. Across participants, as marijuana use shifts from 0 (no) to 1 (yes), the belief that marijuana is relaxing tends to shift from a value of 0 (disagree) to 1 (agree). This association produces a phi correlation of .67. All three research scenarios examine the correlation of marijuana use and the belief that marijuana is relaxing. However, the choice of the appropriate correlation coefficient – Pearson correlation, point-biserial correlation, or phi correlation – depends on whether the variables were assessed using quantitative or binary response options.

Coefficient of Determination

Squaring the correlation coefficient offers a more intuitive interpretation of the relation between two variables. The **coefficient of determination**, or squared value of the correlation (Pearson correlation, point-biserial correlation, or phi correlation), represents the proportion of variance shared between two variables. The higher the value, the greater the amount of variation in one measure that is accounted for by variation in the other, and vice versa. When $r^2 = 1.00$, the proportion of common or shared variation is 100%; and when $r^2 = .00$, the proportion of common variation is 0%. For example, a correlation of .60 (or $-.60$) would yield 36%. This result indicates that scores on the X variable account for or explain 36% of the variation in scores on the Y variable (or vice versa). Or, it may be interpreted that these two variables share 36% of the common variance. The existence of shared variance between two variables does not necessarily indicate whether one variable causes the other.

Regression with a Single Predictor

The scatter plots illustrating positive, negative, and zero correlations depicted in Figure 10.1 do not include a straight line that approximates the linear relation underlying the raw data.

Participant	X	Y
1	0	1
2	0	1
3	0	1
4	0	2
5	1	2
6	2	3
7	2	6
8	6	5
9	8	7
10	9	6

X = "Number of times used marijuana"

Y = "Marijuana is relaxing": 1 (*strongly disagree*) to 7 (*strongly agree*)

Pearson correlation = .86

(a)

Participant	X	Y
1	0	1
2	0	1
3	0	1
4	0	2
5	1	2
6	1	3
7	1	6
8	1	5
9	1	7
10	1	6

X = "Used marijuana or not": 0 (*no*) or 1 (*yes*)

Y = "Marijuana is relaxing": 1 (*strongly disagree*) to 7 (*strongly agree*)

Point-biserial correlation = .78

(b)

Participant	X	Y
1	0	0
2	0	0
3	0	0
4	0	0
5	1	0
6	1	0
7	1	1
8	1	1
9	1	1
10	1	1

X = "Used marijuana or not": 0 (*no*) or 1 (*yes*)

Y = "Marijuana is relaxing or not": 0 (*disagree*) or 1 (*agree*)

Phi correlation = .67

(c)

FIGURE 10.5 Examples of the Pearson, point-biserial, and phi correlations.

Drawing such a line would offer a visual aid to help understand the direction and strength of the relation. It is possible to use the data to mathematically estimate a **regression equation** or model and then apply that equation to draw a regression line or slope that best fits over the observed points of the scatter plot. Regression analysis can be used to enter actual values on one variable (the predictor) for the purpose of forecasting values on another variable (outcome). The best fitting straight line drawn through a scatterplot is called a regression line (see Figure 10.6). The unstandardized regression formula for a regression analysis involving a single predictor is

$$Y' = B_0 + B(X)$$

Y': Prediction for outcome variable score (upon solving the formula).

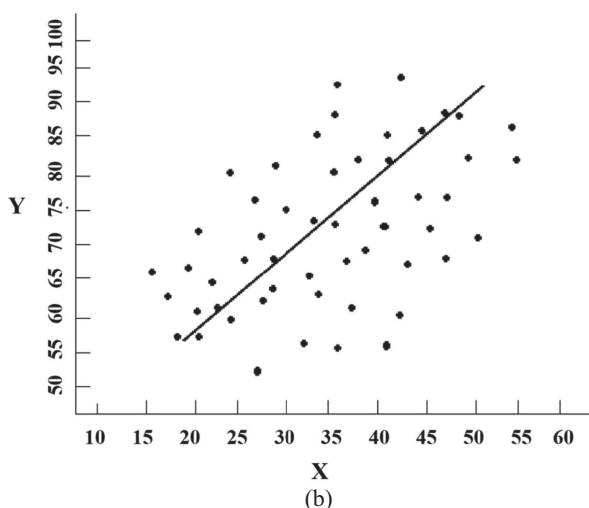
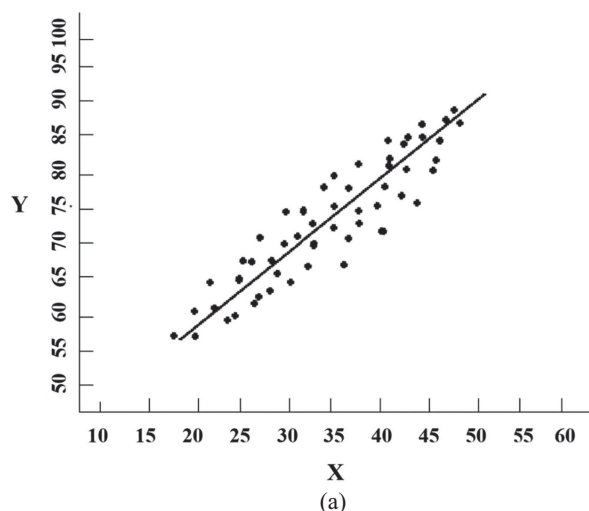


FIGURE 10.6 Scatter plots with regression lines.

B_0 : The constant (or intercept), which provides information on what the predicted Y' score would be if a person scored a zero on variable X .

B : This represents the unstandardized coefficient or slope of the predictor on the outcome. A positive sign indicates a positive association, a negative sign indicates an inverse association, and a value of zero indicates no association between the predictor and outcome. According to the slope value, for every one unit increase in predictor X , a “ B ” unit change in outcome Y' is predicted by the regression formula.

X : Raw score on the predictor variable (to be plugged into the formula).

For example, suppose we know that number of hours of studying predicts scores on an upcoming research methods exam (scored on a 100-point scale). Using data collected from students who took the class last semester, we have developed a regression equation that attempts to predict or forecast performance on the exam for future students who will be taking this course. Based on calculations using the data found in the scatter plot (Figure 10.6a), the regression

formula is as follows: $Y' = 35 + 1.1(X)$. This regression equation is used to derive the regression line plotted in Figure 10.6a.

In this regression equation, Y' is the predicted research methods exam score, and X is the number of hours spent studying. The intercept is 35, with the interpretation that if a student were to study for exactly zero hours, the regression equation would predict or forecast that the student will end up with a poor score of 35 on the exam. Math: $Y' = 35 + 1.1(0) = 35$. The B coefficient or slope shows a positive value of 1.1, and therefore signifies that a positive relation is exhibited between the predictor and outcome. Technically, this B coefficient value could be interpreted more precisely as follows: For every one hour increase in hours studying (X), the regression formula is predicting or forecasting a 1.1 increase in points earned on the exam (Y'). For example, a five-hour increase in the number of hours studying would correspond to a predicted 5.5 point increase on the exam. Math: $1.1(5) = 5.5$. Furthermore, a p-value could be calculated to determine if this B coefficient of 1.1 is significantly different ($p < .05$) from the null hypothesis B value of 0. (Let us assume that this B coefficient is statistically significant beyond chance.)

The regression equation is used to plug in various X values to compute predicted scores on Y (Y'). For example, based on the regression equation and plotted regression line, a mediocre student who spent only 20 hours studying would be expected to obtain a failing score of 57 on the exam. Math: $Y' = 35 + 1.1(20) = 57$. However, a conscientious student who spent 50 hours studying would be expected to earn a much higher exam score of 90. Math: $Y' = 35 + 1.1(50) = 90$.

The Y' or predicted criterion scores obtained based on solving for various X values of the regression formula are predicted values: They do not always correspond exactly to the actual Y scores for all individuals with a specific X value. This is depicted in the discrepancies between the observed scatter plot points and the plotted regression line as illustrated in Figure 10.6a. Perfect prediction of the Y' score stemming from knowledge of a participant's X score only occurs if the correlation is exactly +1.00 or -1.00. The degree of *residuals*, defined as the vertical distances of observed scatter plot points (Y) from the predicted points (Y') of the regression line, is known as the residual or prediction error. A relatively strong correlation (e.g., $r = .80$) between variables X and Y will exhibit smaller residuals or prediction errors and observed data points will be closer to the regression line. A weak correlation (e.g., $r = .10$), on the other hand, will have considerable residuals or prediction error, and data points will be scattered farther away from the regression line.

Figure 10.6 illustrates a comparison between two scatter plots involving measures that are represented by the same regression line derived from the same regression equation intercept and B coefficient but differing in the magnitude of the residuals. In the case of Figure 10.6a, the deviations involving the actual scatter plot points from the regression line are small and so this regression model will yield a small residual value. Thus, for any person, the deviation between their actual score (Y) versus the predicted (Y') score is relatively small. In Figure 10.6b, on the other hand, for each person, the deviation between the scatter plot point (Y) and predicted score (Y') is relatively large and consequently this regression model will yield a larger residual value. Thus, it is always desirable to have a small residual value for the regression model, as that reflects the accuracy of the regression line (as derived from the regression formula) in corresponding to the observed data points. In other words, with a smaller residual value for the regression model, the X variable is more accurate in predicting the Y variable. The residual is inversely related to the magnitude of a correlation coefficient. A smaller residual will produce a larger correlation coefficient. A residual of exactly 0, suggesting that all the scatter points fall exactly on the regression line, will yield a correlation coefficient of 1.00 or -1.00.

Multiple Regression

Just as the experimental designs discussed in Chapter 5 were not limited to the manipulation of a single independent variable, correlational analyses are not limited to the investigation of a single predictor variable. **Multiple regression** is an extension of the bivariate correlation and simple regression to estimate relations involving two or more predictors on an outcome (criterion). A researcher may wish to know in what way the combination of several different predictor variables is related to an outcome measure. For example, we may be interested in how well predictors such as years of education, socioeconomic status, and openness to experience each uniquely helps explain the number of friends from other racial groups. For such an endeavor, the most used analytic technique is multiple regression, in which a weighted combination of predictor variables is used to estimate predicted outcome values on the criterion variable.

Unstandardized Regression Equation

A multiple regression line may be estimated using the following unstandardized multiple regression equation:

$$Y' = B_0 + B_1(X_1) + B_2(X_2) + B_3(X_3) \dots$$

This is very similar to the regression equation for a single predictor, except that the equation for multiple regression allows for the estimation and statistical control of more than one predictor. Let's say X_1 indicates the number of years of education, X_2 is socioeconomic status, X_3 is openness to experience, and Y' is the predicted number of cross-race friends. The intercept or constant (B_0) represents the predicted number of friends if a person scored a zero on all three predictor variables. Each B coefficient represents the unique slope or weight for each predictor variable, while statistically controlling for the competing effects of all other entered predictors on the outcome. Essentially, each B value provides information regarding the unique contribution of each predictor over and beyond all other predictors estimated in the analysis. Thus, B_1 is the unique slope coefficient for predictor X_1 , B_2 is the unique slope coefficient for predictor X_2 , and B_3 is the unique slope coefficient for predictor X_3 . A regression analysis containing more than three predictors would entail additional B coefficients.

As calculated from a sample of participants, the multiple regression equation to predict Y' scores might be as follows: $Y' = 2 + 0.5(X_1) + 1.5(X_2) + 2.8(X_3)$. To obtain the predicted number of cross-race friends, we would need to enter a person's scores for years of education (X_1), socioeconomic status (X_2), and openness to experience (X_3). The intercept value is interpreted as follows. If a person scores exactly a 0 on each of the predictors of years of education, socioeconomic status, and openness to experience, the regression equation will predict or forecast that the person will have two friends from other racial groups. Math: $Y' = 2 + 0.5(0) + 1.5(0) + 2.8(0) = 2$.

The sign of the B coefficient for each predictor provides information about the direction of the relation for each predictor on the outcome, while statistically controlling for all other predictors in the analysis. The B coefficients for all the predictors are all positive. Thus, this indicates that higher years of education, higher socioeconomic status, and higher openness to experience each uniquely contribute to greater number of friends from other racial groups. Specifically, the value of the B coefficient signifies the unique slope effect on the outcome, after controlling for other predictors in the equation. For example, the $B = 0.5$ for years of education indicates that for every one unit increase in number of years of education, the regression equation predicts an

increase in 0.5 number of cross-race friends (after controlling for the other estimated predictors). Thus, a person with two additional years of education would be predicted, by the regression model, to have one additional cross-race friend ($0.5(2) = 1.0$). Furthermore, it is also possible to determine whether the B coefficient for each predictor is significantly different ($p < .05$) than a null hypothesis $B = 0$ of no relation.

Standardized Regression Equation

So far, the interpretation of simple regression and multiple regression has focused on the unstandardized equations involving B coefficients. The unstandardized regression equation is calculated based on the raw (unstandardized) scores for the variables, regardless of their original measurement scales. This refers to the fact that the variables entered into regression analysis could each be assessed using different measurement metrics, such as a seven-point or 100-point scale. As each predictor could be measured using different types of response options, the B coefficients for the different predictors from the same regression model cannot be compared to one another. The unstandardized regression equation presented in an earlier example was as follows: $Y' = 2 + 0.5(X_1) + 1.5(X_2) + 2.8(X_3)$. As such, it would be incorrect to compare the B coefficients and conclude that the predictor of openness to experience ($B_3 = 2.8$) is the strongest predictor and the number of years of education ($B_1 = 0.5$) is the weakest predictor of the number of cross-race friends. Because these different predictors were assessed using different measurement scales, it is not meaningful to compare the B (unstandardized regression) coefficients of different predictors in terms of strength of relation in contributing to the outcome.

If the coefficients are to be compared, an unstandardized regression equation involving B coefficients must be converted to a standardized regression equation involving β ("beta") coefficients. Mathematically, the standardized regression equation is computed as follows. First, all the variables to be used in regression analysis are standardized via the approach known as z scoring. The **z score** formula is presented below:

$$z \text{ score} = \frac{X - M}{SD}$$

For each participant, a subtraction is made between each person's raw score (X) and the mean score (M) and then divided by the standard deviation (SD) for the sample. This transformation is performed for every variable (predictors and outcome) that will eventually be entered into the regression analysis. Inspecting the z score formula, a participant who is scoring above the mean would have a positive z score, a participant who is scoring at exactly at the mean would have a z score of exactly zero, and a participant who is scoring below the mean would have a negative z score. The purpose of z scoring is to standardize all the variables so that they now assume this new common metric. Specifically, after z scoring a variable, the transformed version of that variable will now have a mean of 0 and a standard deviation of 1. Standardization makes the variables "scale free," referring to the fact that the original variables have been released from the constraints of their original measurement metrics. Afterward, the transformed z scored variables are used to estimate and derive the standardized regression equation, presented below:

$$z'_Y = \beta_1(z_{X_1}) + \beta_2(z_{X_2}) + \beta_3(z_{X_3}) \dots$$

This standardized multiple regression equation immediately above may be contrasted with the unstandardized multiple regression equation presented earlier in the chapter. First, notice that the unstandardized regression formula contains a constant (intercept), but the standardized

regression formula no longer contains a constant. Due to the z scoring process, in which the mean of a z scored variable becomes zero, the constant of a standardized regression equation becomes zero and therefore is no longer represented in the formula. Second, instead of B coefficients in an unstandardized regression equation, the standardized regression equation involves β coefficients (standardized coefficients). Owing to the z scoring process in transforming all the variables into the same standardized metric, the β coefficients of the different predictors may now be compared in terms of strength in explicating the outcome.

The values for β typically range from -1.00 to $+1.00$. A positive value indicates a positive unique relation, a negative value indicates a unique inverse relation, and a zero value represents no relation between the predictor and outcome, after controlling for all other predictors in the model. The β coefficient is interpreted approximately like a correlation, but one that has statistically controlled for all other predictors in the analysis. In fact, a standardized regression equation containing exactly one predictor (simple regression) will yield a β coefficient identical to the correlation value between that predictor and the outcome (as no other predictors are controlled for in the analysis).

The standardized regression equation in predicting the number of cross-race friends is written as follows: $z\hat{Y} = .50(z_{X_1}) + .10(z_{X_2}) + .30(z_{X_3})$. Here, in comparing the magnitude of the β weights, years of education ($\beta_1 = .50$) is the strongest predictor and socioeconomic status ($\beta_2 = .10$) is the weakest unique predictor of the number of cross-race friends. Essentially, the β represents an effect size of the unique relation from the predictor to the outcome, while controlling for other predictors. Software programs automatically generate both unstandardized and standardized versions of the regression equation as part of the statistical output. Thus, researchers do not need to z score all the variables before performing regression analysis.

Multiple R^2

The proportion or percent of variation accounted for by the set of predictors on the outcome is noted by the multiple R^2 . The multiple R^2 is an extension of the coefficient of determination (r^2) for more than one predictor. A regression model containing predictors that exhibit larger β coefficients will tend to produce a larger multiple R^2 value. A larger rather than smaller multiple R^2 is desired because it indicates that the set of predictors collectively explains more of the variability in the criterion. The multiple R^2 represents the effect size of all the predictors on the outcome in the regression model and could range from 0 to 1.00, if interpreted as a proportion of the outcome explained (or, alternatively, 0% to 100% of the variance explained). Suppose that the three predictors of the number of years of education, socioeconomic status, and openness to experience collectively explained 35% of the variance in the number of cross-race friends. Although this reveals a considerable amount of variance explained, potential additional predictors might increase this value. Conversely, 65% of the outcome variance is not accounted for or explained by these three predictors. A regression analysis disclosing a low R^2 value is more likely to be considered an *underspecified model*, that is, one in which important and relevant predictors that should have been incorporated to help explain the outcome were not entered into the analysis.

The regular multiple R^2 value, described in the previous paragraph, does not correct for chance. A variation of that index that corrects for chance is referred to as the adjusted (or shrunken) multiple R^2 . For the adjusted multiple R^2 , the value only increases if the entered predictors uniquely contribute to the outcome beyond chance. Consequently, the adjusted R^2 value is always equal to or smaller than the regular R^2 value. “Chance,” as defined by the adjusted R^2 , is attributed to an increase in the number of predictors and a decrease in sample

size. Specifically, a greater discrepancy between the adjusted R^2 compared to regular R^2 will be found when the number of predictors increases. In excessively entering a greater number of unnecessary predictors, some of these variables, by chance, may incrementally increase the outcome variance (as nonrelevant predictors can contribute close to zero or trivial β). Moreover, a greater discrepancy is shown between the adjusted R^2 relative to the regular R^2 as the sample size decreases. Analyses based on a smaller sample size possess greater sampling error and are less representative of the population of interest. Taken together, if the purpose is to yield an adjusted R^2 that is not that much smaller than the regular R^2 , it is best to avoid entering too many unnecessary predictors unrelated to the outcome and to avoid a small sample size. In most published regression analysis studies, the regular R^2 and adjusted R^2 produce relatively similar values. Our regression example involving a relatively small sample size and three predictors to explain number of cross-race friends may yield an adjusted R^2 of 32%, in comparison to the regular R^2 value of 35%.

Experimental Control versus Statistical Control

Perhaps the most important reason for performing multiple regression compared to a correlation or regression with a single predictor is that two or more predictors could be statistically controlled for in the model. That is, the interpretation of a predictor is controlling for or removing the effects of all other entered predictors, to help isolate the unique contribution of each predictor on an outcome. At this point, it is important to elaborate on what is meant by “controlled for” and to distinguish between experimental control versus statistical control. *Experimental control* is pursued in experimental methods, notably in randomized experiments, to ensure that the conditions or groups are initially equivalent so that the impact of the independent variable on the dependent variable can be causally determined. In applying experimental control, the process of random assignment to groups ensures that all possible extraneous variables are equalized across the groups, thanks to chance, and therefore such unmeasured variables could be ruled out as competing explanations on the dependent variable. This permits researchers to make causal claims regarding the effect of the independent variable on the dependent variable. Thus, experimental control is considered the most desirable type of “controlling for” to rule out virtually all possible rival extraneous variables (that don’t even have to be measured by the researcher) that may confound the independent variable to dependent variable relation.

On the other hand, statistical control is pursued in nonexperimental methods to compensate for the fact that experimental control was not implemented. In applying statistical controls, such as estimating a multiple regression analysis containing two or more predictors, the attempt is to rule out rival predictors (covariates) as extraneous variables in the regression analysis. In an earlier multiple regression example, the three predictors of the number of years of education, socioeconomic status, and openness to experience were simultaneously entered and statistically controlled to assess the unique contribution of each on the number of cross-race friends. Keep in mind that additional predictors, such as communication skills and extraversion, if statistically controlled for, may instead reveal that some or all three of the original predictors would no longer uniquely predict the outcome in the regression analysis. Communication skills and extraversion were not measured and not controlled for in the regression analysis, so these extraneous variables may not be ruled out as alternative explanations of the outcome. The greater the number of predictors that are assessed and entered into a regression analysis, the more variables are controlled for and ruled out as competing explanations of the outcome. The major disadvantage of statistical control techniques, such as multiple regression, is that it is nearly impossible to measure and enter all possible variables that could serve as competing explanations on the

outcome. The regression model is only as good as the potential predictors that were explicitly measured by the researcher and therefore statistically controlled for in the model.

When important predictors are not entered and statistically controlled for in the analysis, this is known as model under-specification. Thus, when conducting research to determine whether predictors might be uniquely associated with an outcome, it is strategic to measure as many *relevant* predictors as possible so that they could later be accounted for in multiple regression. The approach can be informed by reviewing the previously published literature on the important predictors and antecedents that may contribute to or operate as rival contributors to the outcome. The greater the number of predictors incorporated, the larger the N needed to get more stable estimates in the regression analysis.

Using Regression to Test Mediation

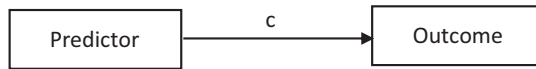
Regression can be applied to test more sophisticated statistical models than just a set of predictors on an outcome. Regression analysis could be used to test mediational models that seek to evaluate the mechanism or process from a predictor to a mediator and then to an outcome. For example, a study might seek to evaluate the pathways from attitudes (predictor) to intentions (mediator) to behaviors (outcome). Essentially, this would evaluate whether the predictive relation from attitudes to behavior is mediated by intentions.

Baron and Kenny (1986) provide guidance for a four-step approach to test and understand mediation. The four steps are presented in Figure 10.7. This approach involves estimating several sets of regression analyses to build an overall mediational model. In Step 1, the purpose is to demonstrate whether the predictor is significantly related to the outcome (path c), not controlling for the mediator. Statistically, perform a regression analysis in which only the predictor is entered to explain the outcome. Conceptually, this step helps to establish that the pathway from the predictor to outcome connection exists (beyond chance), before later taking into consideration the mediator. If this simple predictor to outcome connection is not demonstrated, then it does not make sense to proceed to the other steps.

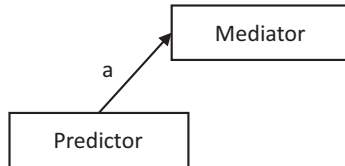
In Step 2, the purpose is to determine whether the predictor is significantly related to the mediator (path a). Statistically, this involves entering the predictor to explain the mediator in a regression analysis. Conceptually, the goal is to demonstrate that the pathway from the predictor to the mediator exists (beyond chance). In Step 3, the purpose is to determine if the mediator is significantly related to the outcome (path b), even after controlling for the predictor (path c'). Statistically, both the predictor and mediator are simultaneously entered as variables to explain the outcome in this regression analysis. Conceptually, this step demonstrates that even after controlling for and therefore taking into account the effect of the predictor, the mediator plays a unique role in explaining the outcome.

If Steps 1 to 3 are not satisfied, the mediational process is not supported. In other words, the predictor to outcome connection is not mediated by this third variable. However, if all three steps are satisfied, this signifies that some extent of mediation (partial or full) is evidenced in the mediational model, so it would be appropriate to proceed with the fourth step. In Step 4, the regression coefficient from the predictor to the outcome (path c) without controlling for the mediator is compared to the regression coefficient from the predictor to outcome (c') while controlling for the mediator. *Partial mediation* is established if the path c' (compared to path c) coefficient is reduced but remains statistically significant. *Full mediation* is established if the path c' (compared to path c) coefficient is reduced to near zero and no longer statistically significant. Thus, full mediation occurs when the mediator, after entry into the regression model, helps to entirely account for or explain the process from the predictor to the outcome.

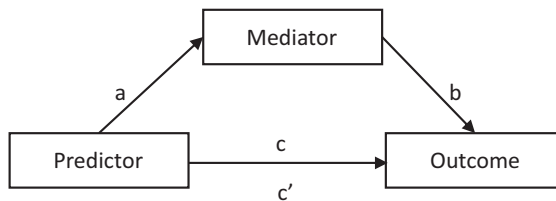
Step 1. The predictor is significantly (path c) related to the outcome (not controlling for mediator).



Step 2. The predictor is significantly (path a) related to the mediator.



Step 3. The mediator is significantly (path b) related to the outcome, after controlling for the predictor (path c').



Step 4. If the Steps 1 to 3 are satisfied, compare paths c vs c'.

- **Partial mediation:** Path c' (vs. c) coefficient is reduced and still significant.
- **Full mediation:** Path c' (vs. c) coefficient is reduced to near 0 and no longer significant.

FIGURE 10.7 Testing mediation with multiple regression using the Baron and Kenny (1986) approach.

Note: “Significantly” and “significant” refer to statistical significance beyond chance, usually tested at $p < .05$.

The Baron and Kenny (1986) four-step approach is widely applied in research studies and provides a step-by-step foundation to understand how to statistically test and interpret mediational models using regression analysis. However, some criticisms have been levied against this four-step approach. First, this approach is sometimes considered too statistically conservative, as some have put forth that Step 4 is the most important step and that testing all four steps is unnecessary. Second, although it is possible to extrapolate the four-step approach beyond three variables to evaluate multiple predictor(s), mediator(s), and outcome(s) in a mediational model, this is unwieldy due to the number of regression analyses that ought to be conducted to build this overall model. A more efficient and sophisticated strategy to test complex mediational mechanisms involving many variables is using structural equation modeling (described in the next chapter).

Uses and Misuses of Correlational Analysis

From a research orientation, the main problem with the evaluation of freely occurring variables is that they usually have natural extraneous variables; that is, the occurrence of the variable of

interest is confounded by the co-occurrence of other (usually unmeasured) factors that naturally accompany it and that were not statistically controlled for in the analysis. To take a simple example of a naturally occurring variable, suppose that a researcher is interested in demonstrating a predicted relation between weather and depressed mood. Specifically, the investigator hypothesized a positive relation between rain and depression; the greater the amount of rain, the greater the score on a depression inventory. However, even if the research confirms the existence of a positive association between the occurrence of rain and the level of depression, the investigator is a long way from identifying rain as the underlying causal factor of depression. Frequently co-occurring with rain are other weather conditions such as low barometric pressure, gray skies, and heavy cloud cover, any of which might provide plausible alternative explanations for the occurrence of psychological depression. Only if these other factors were measured and statistically controlled for by being held constant could rain be isolated as uniquely associated with depression. Still, this association should not be casually interpreted, as other unmeasured extraneous variables could still serve as confounds. Unfortunately, until researchers can bring the occurrence of rain under experimental control, they must be aware of the limitations on the interpretation of nonexperimental data.

Hidden Third Factor

The observed relation between any two variables may be affected by a third source of variation that is accidentally or causally linked with one of the observed variables (discussed in Chapter 2). The rain to depression illustration given above is an example of this potential confound in which the natural covariation between rain and low barometric pressure (or gray skies) may confuse our interpretation of the true nature of the causal relation. For instance, the actual relation may be as depicted in Figure 10.8.

In other instances, a third factor may be a causal variable that accounts for the common variation in both observed variables. Such a “third cause,” in the form of some unknown hidden or lurking factor, was at one time proposed as an alternative explanation for the obtained relation between cigarette smoking and the likelihood of developing lung cancer, as depicted in Figure 10.9. Tobacco companies long argued that people who smoke do so because they have a higher sensation seeking personality than the general population. Rather than cigarettes causing cancer, they suggested that such a personality disposition may instead be the culprit in causing these high-risk individuals to not only use nicotine but also to be more susceptible to cancer. As long as correlational studies of the relation between smoking and disease were limited to the observation of people who are already (through self-selection) smokers or non-smokers, such third-factor explanations could never be completely eliminated. The convergence of results from multiple sources of correlational and experimental data, especially randomized

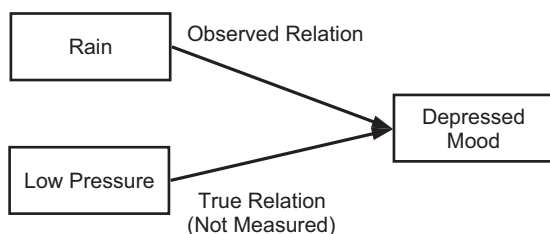


FIGURE 10.8 Hidden third variable.

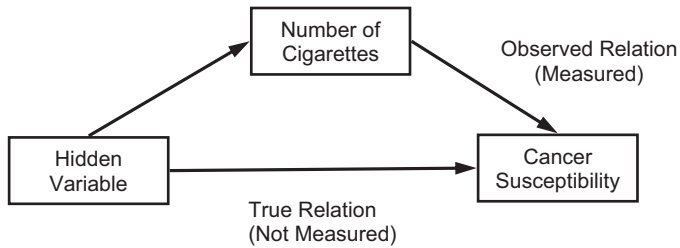


FIGURE 10.9 Hidden variable explanation of the smoking to cancer relation.

experiments involving the effects of cigarette smoke on animals, has convincingly demonstrated that nicotine indeed causes cancer.

Because an association between any two (or more) nonexperimental variables can have multiple possible interpretations, students are usually carefully taught to recognize that “correlation does not necessarily imply causation.” Just because values on variable A can help predict values on variable B does not necessarily mean that A causes B. Of course, a correlational association *can* reflect a causal relation, particularly if one variable precedes the other in time (but not always even then). To be clear, causation implies correlation, but correlation does not (necessarily) imply causation.

Conclusions

Determining the direction and strength of the relation between a specific predictor and a criterion is particularly difficult if the observed predictor variable is only one of several interrelated factors that contribute to variation in the criterion variable. Because many of the covariates that confound naturally occurring phenomena cannot be extricated or made independent of one another, researchers have been tempted to substitute correlational strategies for experimental control in attempts to sort out the factors that contribute to variation in measures of interest. Many of these attempts reflect an inadequate understanding of the context of testing and measurement within which most of our commonly used correlational techniques were developed. Measurement theorists are clear that there is a distinction between a measuring instrument and the underlying construct it is designed to represent (Chapter 4). Any single measurement is usually a partial and inadequate reflection of the underlying variations on the dimension or construct of interest about a particular sample of participants. As discussed, the observed score obtained on any measurement consists of both “true score” and “error” components. The true score represents the reliable, relatively stable characteristics of the feature or event being measured, and measurement error represents the unmeasured systematic or random factors that affect responses at the time of measurement.

Knowledge of the various sources of true score and error components of any measuring instrument is relatively unimportant if one is interested only in the development of prediction equations. Prediction is neither sufficient nor equivalent to a theory or theoretical explanation that helps to deeply understand the predictive relations. If scores on one measure are consistently related to scores on another, then the former may be used to predict the latter. Thus, for instance, scores on a high school achievement test may be applied to predict academic performance, but an uninformed researcher may not truly understand whether the source of common variation is due to extraneous variables such as basic intelligence, motivation, study habits, parental income, or ability to cheat on tests! An appropriate conceptual explanation requires a

comprehensive theoretical understanding of the relations among variables as failing to distinguish between prediction and theoretical explanation can lead to misinterpretation of the results from correlational studies.

References

- Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173–1182.
- Suedfeld, P., Tetlock, P. E., & Streufert, S. (1992). Conceptual/integrative complexity. In C. P. Smith, J. W. Atkinson, D. C. McClelland, & J. Veroff (Eds.), *Motivation and personality: Handbook of thematic content analysis* (pp. 393–400). New York: Cambridge University Press.

ADVANCED MULTIVARIATE CORRELATIONAL DESIGN

The previous chapter described the concepts for interpreting and understanding basic correlational designs involving the Pearson correlation, simple regression, and multiple regression. This current chapter continues with the discussion of correlational designs by focusing on approaches that permit the estimation of more complex relations. The focus will be on common multivariable and multivariate correlational designs: multilevel models, structural equation models, cross-lagged panel models, and growth curve models. Multilevel models are an extension of multiple regression that permits the evaluation of predictors on an outcome when data are collected based on a nested hierarchy. Structural equation models are an extension of multiple regression that permits the estimation of multiple predictors and multiple outcomes in a single analysis. Structural equation modeling has two subtypes: path analysis, which does not use latent factors, and latent structural equation models, which involve latent factors. Cross-lagged panel models are used to unravel the direction of the temporal relations between variables using longitudinal data. Growth curve models are applied to evaluate and identify temporal trends based on longitudinal data.

Multilevel Models

As an extension of multiple regression, a **multilevel model** is used to estimate the relations of predictor(s) to a criterion if the design involves a nested hierarchy of units (Heck & Thomas, 2022; Heck et al., 2013; Kahn, 2011). Multilevel modeling has been used interchangeably with the terms *multilevel regression modeling* and *hierarchical linear modeling*. Multiple regression is appropriate if data are collected at only one level, but multilevel modeling is the appropriate choice if the design is nested with at least two hierarchical levels.

Suppose you conducted a study at a particular bar near the vicinity of your university. Predictor measures were each patron's attitudes toward alcohol and the number of minutes spent at the establishment, and the outcome measure was the number of drinks each patron consumed that night. In this scenario, if the measures were administered to participants found in only one bar, multiple regression is entirely appropriate to estimate the contribution of two predictors (attitudes and time spent) on the criterion (number of drinks consumed). However, a different approach is required if you expand the research geographically to collect measures from patrons

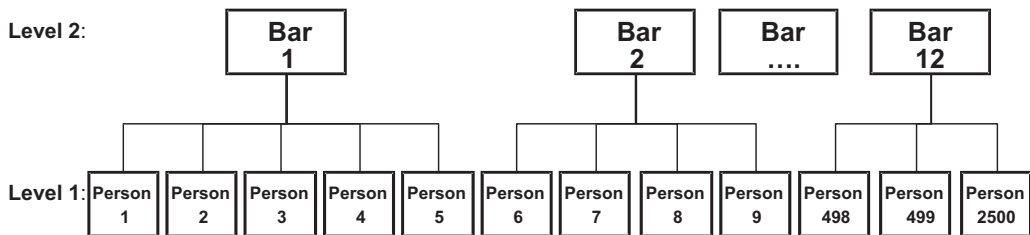


FIGURE 11.1 Multilevel modeling with two levels: 12 bars and 2500 customers.

at multiple bars. A multilevel design is illustrated in Figure 11.1, with data from 2500 customers (Level 1) collected across 12 neighborhood bars (Level 2). Notice that each customer, representing the nested participant unit, is observed in one and only one of the bars. In other words, the design is hierarchical because participants are *nested* within each drinking location. If multiple regression was used to assess the relations of the predictors on the outcome, it would rely on the erroneous assumption that customers were all drinking in the same establishment and that the outcome measure was unaffected by this environment.

This assumption is probably incorrect as it is likely that the amount of drinking (and the factors inducing drinking) varies from bar to bar. In that case, the customers in one bar would be more like each other than they are like customers in a different bar. The **intraclass correlation (ICC)** is an index to assess the extent that the proportion of variance in the outcome is attributed to the grouping structure in a multilevel model. The formula for the ICC below shows that it represents the variance in the outcome scores due to the between-group variance relative to the total variance, which consists of the between-group and within-group variance. In a two-level multilevel model, the between-group variance refers to variance in the outcome measure that is explained by the Level 2 grouping units. The total variability refers to all the variance in the outcome measure explained by the Level 2 grouping units (between-group variance) as well as that explained by the Level 1 participant differences (within-group variance).

Logically, the total variability in the outcome measure could be attributed only to these two sources. Applied to the research scenario depicted in Figure 11.1, the ICC reflects the proportion of variance in number of drinks consumed that is due to mean differences across the bars (relative to the total variability on the outcome measure). As the ICC indicates the degree that scores on the criterion variable differ as a function of the grouping variable, the greater the mean differences across the bars in the number of drinks consumed, the higher the ICC. The ICC ranges from a value of .00 to 1.00, with a value significantly greater than .00 indicating that the clustering or grouping effect due to the nested design cannot be safely ignored. The ICC should not be overlooked because people clustered or nested in the same grouping unit tend to be more similar to one another (e.g., demographically, attitudinally, behaviorally) than they are to people who are nested within a different grouping unit.

$$\text{ICC} = \frac{\sigma_{\text{between-groups}}^2}{\sigma_{\text{between-groups}}^2 + \sigma_{\text{within-groups}}^2}$$

In Figure 11.1, the greater the differences across the different bars in the criterion (the average number of drinks consumed by their respective customers), the higher the ICC. This is because of self-selection biases in which patrons decide to consume alcohol at a particular bar over other bars for various reasons. For example, customers who self-select by choosing to consume their

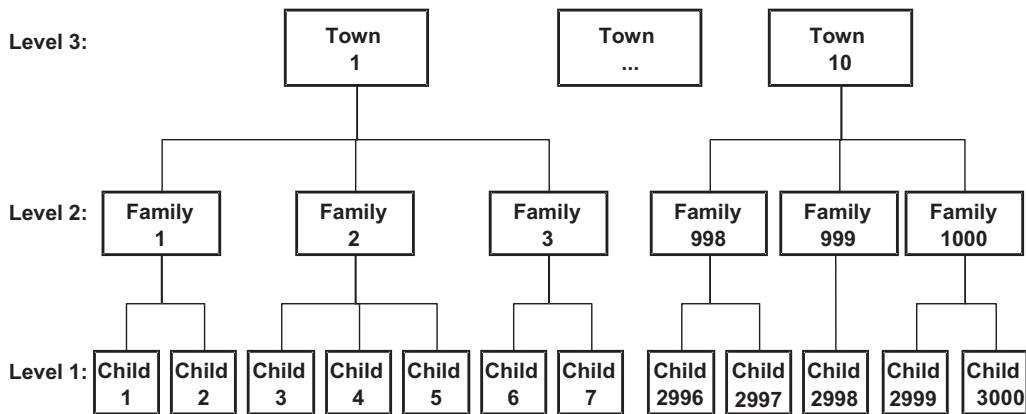


FIGURE 11.2 Multilevel modeling with three levels: ten towns, 1,000 families, and 3,000 children.

alcoholic beverages at a country and western bar might be more similar to (and their alcohol use correlated with) other customers of the same bar than those who drink at a karaoke or hip hop bar. A country and western bar might tend to have customers who are older and more politically conservative, to name some possibilities that may play a role in the number of drinks consumed. These differences across bars may be responsible for confounding the relations involving the hypothesized predictors and the outcome variable. Multilevel modeling statistically controls for, and therefore rules out, the hierarchical design artifact of the ICC, and this enables more accurate estimation of the statistical relations from the predictors to the criterion.

Multilevel modeling research can be undertaken with more than two hierarchical levels. An example of a three-level model is presented in Figure 11.2. Suppose a research project involved data collected from ten towns (level 3) involving 1,000 families (level 2) and 3,000 child participants (level 1) in California. By applying a three-level multilevel model, the design accounts for the possibility that siblings within the same family are more similar to each other (e.g., genetically, demographically, attitudinally, and behaviorally) than to children from other families. Furthermore, a family tends to be more like other families located in the same town across a variety of characteristics (e.g., socioeconomic status and quality of schools) than to families located in different towns.

Aside from allowing the design to estimate the value of the ICC and therefore control for the nested structure, multilevel modeling enables the researcher to evaluate the unique contribution of predictors at different hierarchical levels on an outcome. For example, the design in Figure 11.2 might involve examining predictors that help explain the outcome of children's scores on a verbal ability test. The Level 1 predictors (child level) might be the age of the child and number of books the child typically reads, which can be assessed simultaneously with the Level 2 predictors (family level) of number of books available in each home and each mother's educational attainment and the Level 3 predictor of number of libraries in each town. In this multilevel design, we have a total of five predictors (from three different levels) that are simultaneously entered and controlled for to determine the unique contribution of each in explaining the outcome of children's verbal skills.

The design of Figure 11.2 could be expanded hierarchically, for example, if we receive additional funding to extend this line of research across states (e.g., California, Colorado, Michigan, etc.). This rather complex model consisting of four hierarchical levels would acknowledge and control for the fact that towns located or clustered within the same state tend to be more similar

than towns located in another state on characteristics such as state policies about reading or educational guidelines. Multilevel modeling allows for considerably greater precision in identifying different sources of variation in nested research design. At least ten or more grouping units (e.g., bars, families, towns, or states) at the second level or higher are needed for a multilevel model.

Structural Equation Models

More sophisticated multivariate analytic approaches have become available to evaluate multiple variables in a theory-testing framework. Multiple regression was developed to extend the utility of the Pearson correlation by allowing the inclusion of many predictors, but this extension was restricted to a single criterion outcome. A **structural equation model** overcomes the limitations of multiple regression analysis by allowing researchers to estimate relations among multiple predictors and multiple criterion variables in a single overall analysis. Although advances allow for the integration of structural equation modeling and multilevel modeling, this combined technique is uncommon in the literature due to the complexity of these models. Thus, our elaboration of structural equation modeling focuses on the more widespread non-nested single-level designs.

Structural equation modeling requires the researcher to hypothesize a set of relations that exist among a set of variables, usually based on some theoretical framework. These models require specifications of a set of associations that link theory-relevant constructs and typically are closed systems. A structural equation model, such as correlation and multiple regression, is not necessarily designed to describe the total universe of all the variables that might be related to the variables in the model.

Structural equation models involve understanding diagrammatic terminology, as presented in Figure 11.3. One reason for having a dedicated language is that predictor variables may explain other predictor variables, and outcome variables may, in turn, explain other outcomes. The notation system serves to avoid this tangled Web and offers a standard vocabulary for understanding the variety of possible variables and relations among variables. Structural equation models may be classified into two major types: path models and latent structural equation models.

Path Models

A **path model** is a type of structural equation model in which predictive relations among measured variables are estimated. A variable in a path model may be represented with a single item or by computing an average composite of a set of highly related items. A composite intended to capture a construct should be internally consistent (e.g., Cronbach's alpha) before they are entered in the path analysis model.

Continuing to apply the notation system in Figure 11.3, three path models are shown in Figure 11.4. Similarity is considered an **exogenous variable** (no one-headed arrow is pointing at it) – that is, the variable is not explained by another determinant or predictor postulated in the model. Thus, an exogenous variable operates only as a predictor in the model. Assumed reciprocity and attraction are each classified as an **endogenous variable** (one-headed arrow is pointing at it) – variables explained by a determinant or predictor, as postulated in the structural equation model. An endogenous variable operates as an outcome only or as both a predictor and an outcome in the model.

The E in each path model indicates predictive error – the proportion of variance that was not explained by other variables pointing to it in the model. That is, other variables not measured in

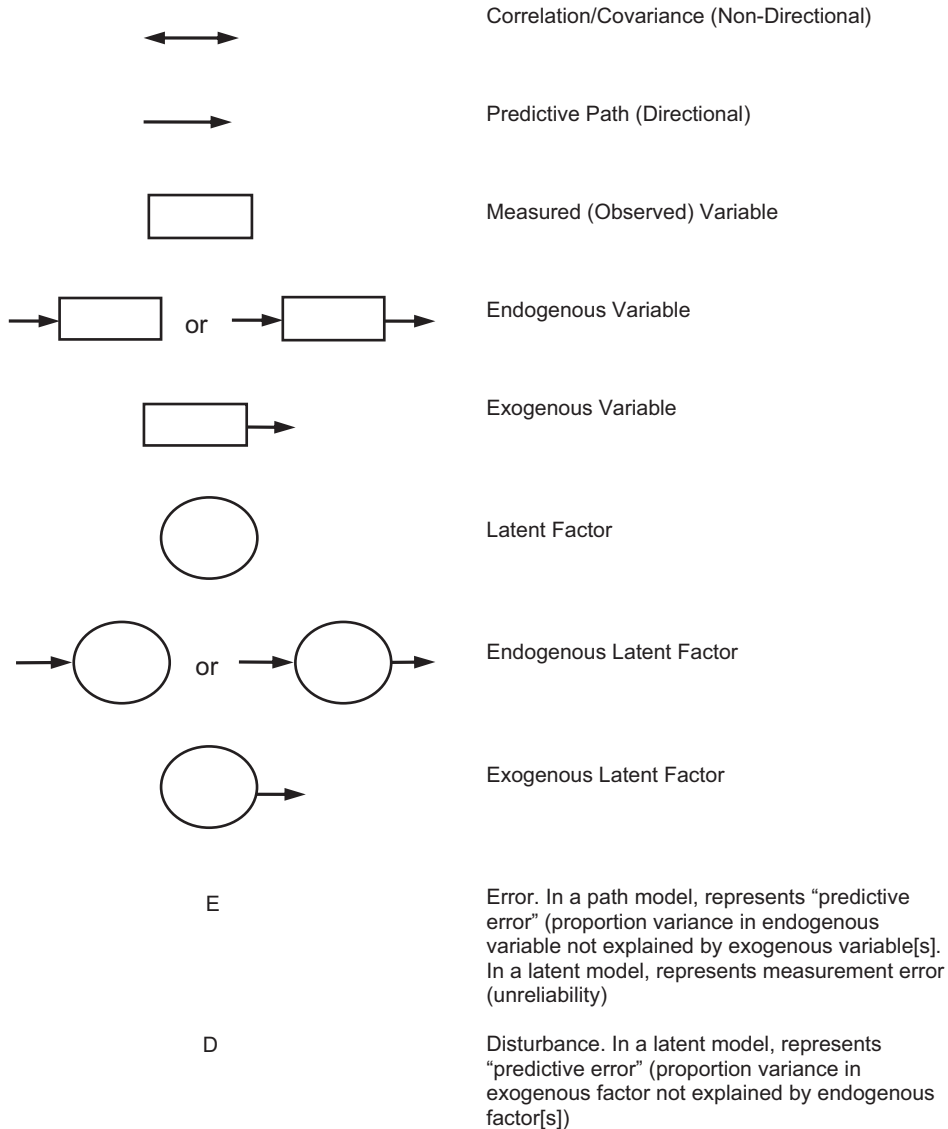


FIGURE 11.3 Notations in a structural equation model.

the investigation may also play a role in contributing to these endogenous variables. Generally, when designing a path model, researchers use the notation of a predictor error to recognize that the model has not specified or measured all the possible variables that might explain the endogenous variables. In Figure 11.4b, for example, the model has specified similarity as the sole predictor of assumed reciprocity, but it is probably not the only determinant, because additional variables (outside the model) also may explain this endogenous variable. The predictive error term indicates our explicit recognition of this fact. This predictive error term in path models is analogous to the residual or predictive error in multiple regression ($E = 1 - R^2$) as discussed in the previous chapter. Thus, a predictive error reflects all the unmeasured determinants of an endogenous variable (including random error).

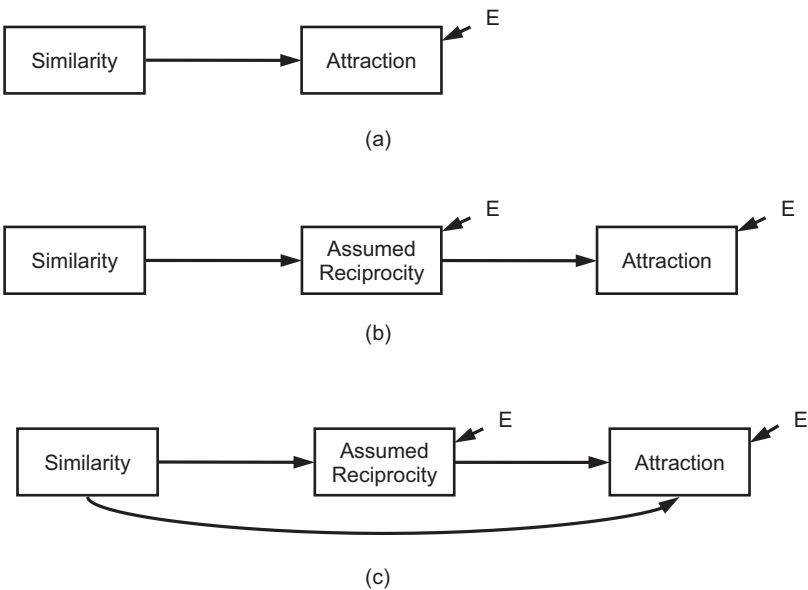


FIGURE 11.4 Path model of hypothesized relations between attitude similarity and attraction.

Mediation. Although mediational processes can be tested with multiple regression (Baron & Kenny, 1986), structural equation models are increasingly being used to analyze mediational hypotheses (Zhao et al., 2010). An advantage of the structural equation modeling approach, including path models and latent predictive models, is that all the variable relations can be estimated in a single analysis, whereas the regression approach requires estimating several regression analyses to build a mediational model (as described in Chapter 10). In testing mediational models, a distinction is made between *direct* and *indirect* effects. With direct effects, one variable directly predicts another variable. Conversely, some variables are thought to explain other variables only indirectly – that is, their influence on a variable is mediated by another variable (or set of variables), as discussed in Chapter 2. These indirect effects are also known as mediated effects.

Figure 11.4 draws on an example from the literature on interpersonal attraction that will help illustrate the distinction between direct and indirect (mediated) paths. In his book *The Attraction Paradigm*, Byrne (1971) argued that attitude similarity causes attraction – that is, we tend to like those who have attitudes similar to ours. To demonstrate this, Byrne manipulated the extent to which participants’ attitudes were shown to be consistent with those of a hypothetical other person. When participants asked if they would like to be friends with and work with this other person, participants in the high attitudinal similarity condition were much more likely to report favorable responses than those in the low attitudinal similarity condition. This finding supporting Byrne’s hypothesis of a direct effect between similarity and attraction is illustrated in Figure 11.4a.

Later research (Condon & Crano, 1988; Napolitan & Goethals, 1979) suggested that the relation demonstrated by Byrne was mediated by another variable, namely, the assumption of reciprocity. Diagrammatically, this fully mediated process is summarized as a path model in Figure 11.4b. This proposed mediation was based on the idea that we tend to like others who are attitudinally similar to us because we assume they will like us back. In this model, two direct effects are hypothesized: similarity to assumed reciprocity and assumed reciprocity to attraction. One indirect effect (or mediational effect) is hypothesized: similarity to assumed reciprocity to

attraction. In other words, the relation between similarity and attraction is completely mediated by the assumption of reciprocated liking. Notice that no arrow is directly pointing from similarity to attraction in this *fully mediated model*. In a fully mediated model, the mediator is entirely sufficient to account for the connection from the predictor to the outcome.

Figure 11.4c represents a *partially mediated model* in which assumed reciprocity operates as a mediator, but in addition, there remains a direct effect of assumed similarity on attraction. In a partially mediated model, the mediator is not entirely sufficient to account for the connection from the predictor to the outcome. In Figure 11.4c, the three direct effects are similarity to assumed reciprocity, assumed reciprocity to attraction, and similarity to attraction. An indirect effect occurs starting with similarity through the mediator of assumed reciprocity to attraction. The traversal of pathways via indirect effects may be determined by starting the trace (through one-headed arrows) from an initial exogenous variable to a mediator to a final endogenous variable.

To test the fully mediated model (Figure 11.4b) and partially mediated model (Figure 11.4c), Condon and Crano (1988) used Byrne's (1971) method for manipulating assumed similarity. Participants were asked to rate (1) how similar they were to the other person (similarity), (2) the extent to which they believed the other person liked them (assumed reciprocity), and (3) how much they thought they would like the other person (attraction). The mediational model could be tested with path analysis. The path from similarity to assumed reciprocity was statistically significant ($p < .05$), as was the path from assumed reciprocity to attraction. If only these paths were statistically significant, support would be provided for the fully mediated model in Figure 11.4b. However, the path from similarity to attraction also was significant, even after controlling for the mediating role of assumed reciprocity. Thus, the results showed that assumed reciprocity partially mediated the connection from similarity to attraction and thereby supported the model diagrammed in Figure 11.4c. This analysis suggests that both direct and indirect effects from similarity to attraction were plausible. The mediational interpretation is not contradictory to the direct effect idea; rather, it shows that assumed reciprocity does not fully account for the connection from similarity to attraction. This result encourages us to search for additional potential mediators that could help to fully mediate this predictor to outcome connection.

Multiple mediation. Considerable research demonstrates that we are attracted to good-looking people (Dion et al., 1972). Some have hypothesized that this relation between physical attractiveness (beauty) and liking is mediated by perceived social competence (i.e., physically attractive people are more socially competent and therefore more likable). The path model of Figure 11.5 integrates these two lines of research on attraction and is an example of a model involving multiple mediators (Preacher & Hayes, 2008). In this hypothesized model, two exogenous variables are beauty and similarity, whereas the three endogenous variables are social competence, assumed reciprocity, and attraction. The relations from beauty and similarity to attraction are postulated to be mediated by social competence and assumed reciprocity. Specifically, the pathway from similarity to attraction is hypothesized to be mediated through assumed reciprocity. Furthermore, the pathway from beauty to attraction is hypothesized to be mediated through both social competence and assumed reciprocity.

The rationale for these two mediators is understood as follows. At an early age, good-looking children appear to be more pleasant to interact with than unattractive children (Berry & McArthur, 1986) and consequently receive more attention from parents and other people. When these children become adults, they have already learned more social graces and are more socially competent than unattractive children who do not receive as much positive adult attention. In addition, a socially competent person is assumed to be more likely to reciprocate. Given this

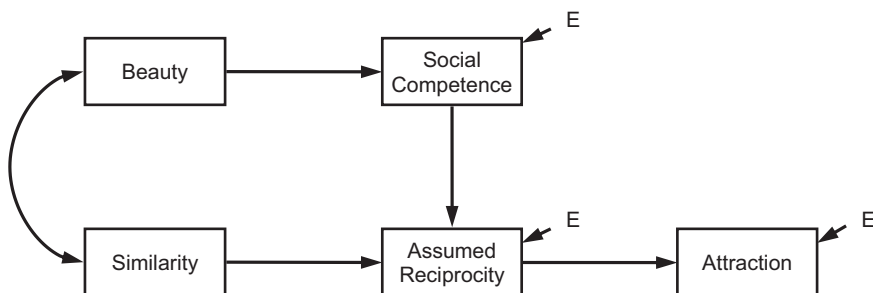


FIGURE 11.5 Path model linking physical beauty and similarity to attraction.

chain of events (Figure 11.5), we hypothesize that people who are physically beautiful tend to be liked (attraction), but this effect is mediated through the two mediating variables of social competence and assumed reciprocity. Based on tracing the pathways of the entire model, four direct effects are hypothesized: beauty to social competence, social competence to assumed reciprocity, assumed reciprocity to attraction, and similarity to assumed reciprocity. An indirect effect is found from beauty to social competence to assumed reciprocity to attraction. An indirect effect also is shown from similarity to assumed reciprocity to attraction.

Observant readers will notice that two different types of arrows (single-headed vs. double-headed) connect the variables in Figure 11.5. Most common in this diagram are the single-headed arrows – for example, those linking social competence and similarity to assumed reciprocity. Connections of this type imply directional hypotheses; thus, similarity and social competence are hypothesized to predict assumed reciprocity uniquely, which in turn is hypothesized to predict attraction. But notice the connection in Figure 11.5 between beauty and similarity. Here, a double-headed arrow connects the variables. A double-headed arrow indicates that a nondirectional bivariate correlational relation is hypothesized between the variables. As was noted, both beauty and similarity are exogenous variables, and the other three variables are endogenous variables. A third type of relation also is possible, namely, no relation. In these instances of variables that are hypothesized to be unrelated, no arrow connects the variables. In Figure 11.5, no direct effect is hypothesized between beauty and assumed reciprocity, indicating that no statistically significant linkage is expected to exist between these variables. This hypothesized mediational model will need to be tested by collecting participant data on these variables and afterward specifying a path analysis to determine if the specified coefficients are indeed statistically significant.

Latent Structural Equation Models

A **latent structural equation model** or latent predictive model is an integration of confirmatory factor analysis and path analysis. All the path analysis issues discussed up to this point also apply to latent structural equation modeling. However, a path model uses only measured variables to represent constructs. This route is taken if the researcher decides not to account for measurement error in each factor. Thus, a weakness of a path model is that the measurement error of each factor is not estimated and therefore not statistically removed in the analysis. This is not the case in latent structural equation models, in which the measurement error is statistically removed from each factor before testing predictive relations across the factors. A latent structural equation model contains a *measurement component* derived from factor loadings in confirmatory factor analysis and a *structural component* representing the predictive relations between

factors. Essentially, each construct in the model has been factor-analyzed so that the variable is “free” from measurement error. This pure factor, uncontaminated by measurement error, is also known as a **latent factor** and is used to specify predictive relations with other exogenous or endogenous latent factors. The notation system given in Figure 11.3 is necessary to understand latent structural equation models. Suppose you are excited and delighted that the circus is coming to your college town. As this traveling circus will be visiting for an entire month, you take advantage of this opportunity to investigate a hypothesized model linking beliefs and behaviors about attending the circus. The Theory of Reasoned Action (Ajzen & Fishbein, 1980) serves as the guiding framework. This theory of human motivation involves five factors: attitudes, social norms, intentions, and behavior. In this theoretical framework, personal attitudes and social norms regarding that behavior are hypothesized as correlated. Both constructs are postulated as simultaneously explaining intentions to engage in the behavior. Intentions, in turn, are postulated to predict the behavior.

In a study designed to test these relations, a questionnaire is created containing 11 questions that are administered to a sample of participants. Participants rate the extent that they personally disagree or agree with each of three attitudinal items regarding the circus being fun, exciting, and enjoyable. Norms are measured with three questions regarding beliefs about how much their best friend, sibling, and mother like the circus. The intentions construct was assessed with three items that ask about the extent that respondents want to, seek to, and plan to visit the circus. A month later, the same participants were asked two behavioral questions concerning the number of days and hours they had spent at the circus.

If estimating these hypothesized connections using the path model route, our set of predictive relations might resemble the diagram depicted in Figure 11.6. First, the four variables are derived by computing the mean of their respective items. The path model is then estimated to determine whether the predictive relations among these measured variables are statistically significant.

Instead, if the decision was to estimate a latent structural equation model, the researcher should first estimate the measurement component using confirmatory factor analysis to determine the association between each individual item and its corresponding factor. Figure 11.7 shows a confirmatory factor analysis involving correlations among the four latent factors, which are tapped by their corresponding measured items. The large circles represent latent factors, called so because these factors are not directly assessed or observed. Rather, the variance of a latent factor is extracted from the shared or common variance of its measured items. For example, the latent factor of attitudes was mathematically derived from the variance that is shared across its three measured items (fun, exciting, and enjoyable). Items that are internally consistent or highly correlated with other items will have more shared variance.

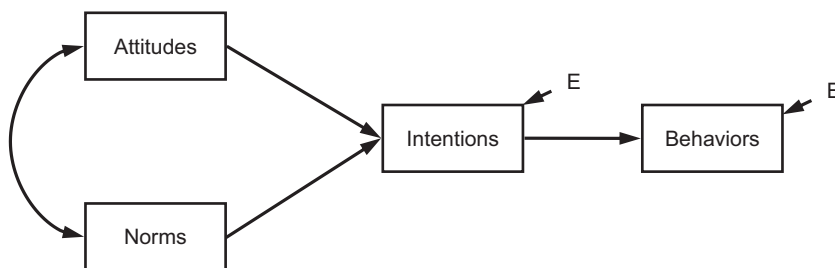


FIGURE 11.6 Path model of theory of reasoned action.

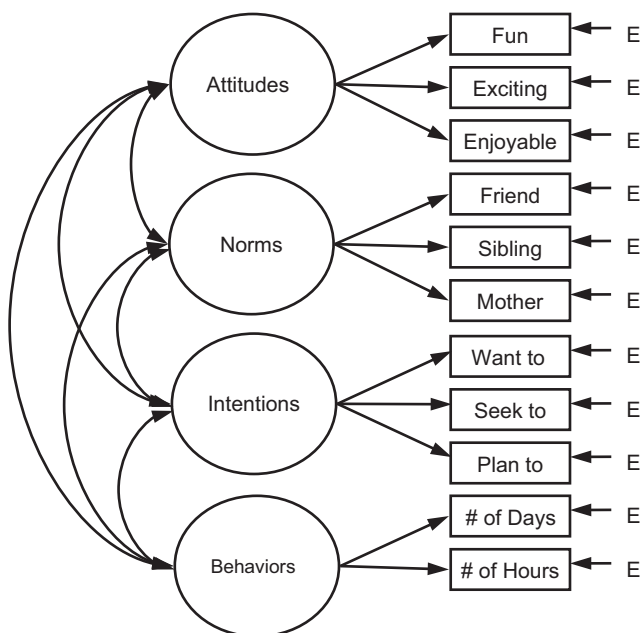


FIGURE 11.7 Confirmatory factor analysis.

The E in a latent model represents measurement error. A measurement error of an item represents the unreliability of that item in not sharing variance with the other items in that factor. A low measurement error value is desired, as it suggests that its corresponding measured item is “hanging together” or internally consistent with the other items in that factor. The path from the latent factor to a measured item represents a factor loading. A factor loading represents how representative the item is of its factor. A factor loading may range from 0 to 1, with a higher compared to lower value denoting that an item better loads on its latent factor. In other words, items that exhibit high factor loadings suggest high internal consistency with the other items on that latent factor. An item exhibiting a low factor loading is undesirable as it shows that the item does not reliably “hang together” with the other items that load on that factor. The nondirectional correlation between two latent factors is signified by an arrow with two heads. However, this is no ordinary Pearson correlation (although interpreted similarly), but a correlation in which measurement error has already been statistically removed from the factors through the factor analytic extraction.

After determining that the measurement component is acceptable as evidenced by high factor loadings on the latent factors, is it then acceptable to interpret the structural component that involves predictive connections among the factors. (After all, if the items in a factor are not reliable measurements, it would be futile to use such a poor factor in any meaningful way in a hypothesized framework.) The latent structural equation model is presented in Figure 11.8. Essentially, the measurement component and structural component are simultaneously estimated in a latent structural equation model. The association between each latent factor and its measured items represents the measurement component, as given in the confirmatory factor analysis earlier. The predictive relations between the various latent factors represent the structural model. The predictive paths between latent factors denote the relation after measurement error has been statistically removed from the factors. This is the reason why latent techniques,

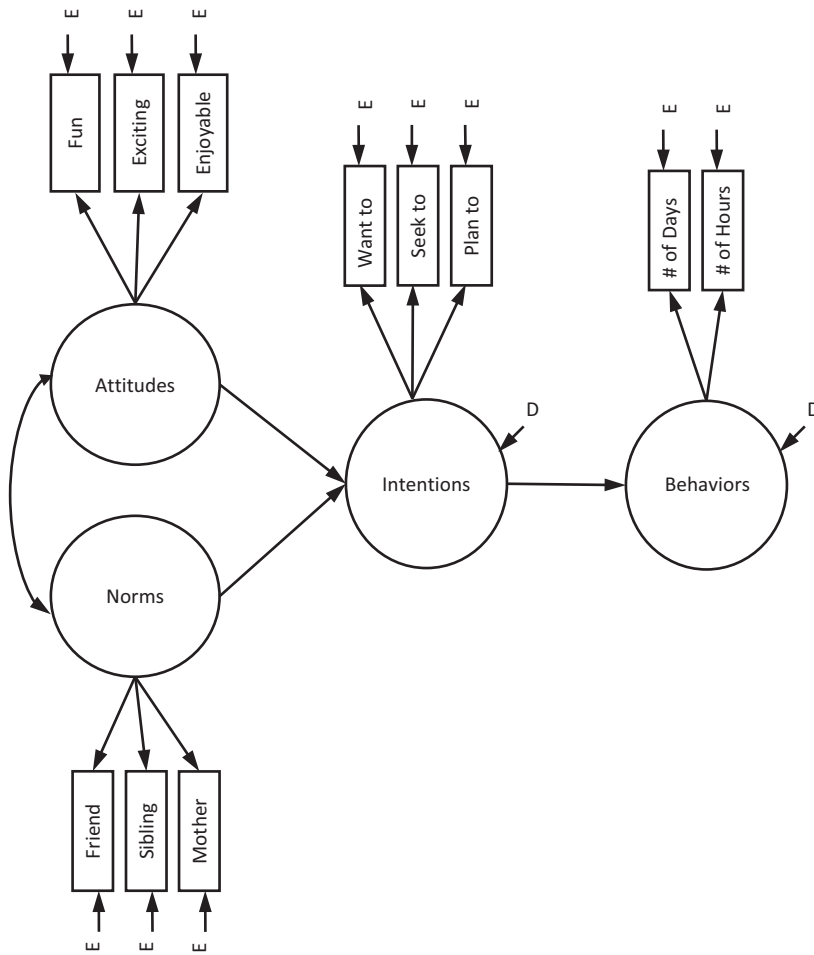


FIGURE 11.8 Latent structural equation model.

like latent structural equation modeling, are sometimes said to be “free” from measurement error. However, this statement is only true to the extent that the measured items are representative of the underlying factor of interest, but it is impossible to know the entire world of all possible measured items that represent a theoretical construct. Thus, it is advantageous to use a variety of highly related items with the goal of fully capturing the conceptual bandwidth of the construct.

Next, the structural component of the latent structural equation model of Figure 11.8 is interpreted. Attitudes and norms serve as the latent exogenous factors as each only has a predictive role (with no other variable or factor pointing at it) in the model. Intentions and behavior are latent endogenous factors that are being explicated or determined by other factors in the model. The “D” in the figure indicates the *disturbance* term, essentially the predictive error (i.e., unexplained variance) stemming from an endogenous latent factor not being fully explained by other factors in the model. The disturbance term denotes that unassessed latent factors outside the contained system shown in the diagram might also influence these endogenous factors. The direct effects are the estimated one-headed paths involving latent factors in the model: (attitudes to intentions, norms to intentions, and intentions to behavior). Tracing the traversal of one-directional pathways, two indirect effects are possible: attitudes to intentions to behaviors, and

norms to intentions to behaviors. Thus, the pathways from attitudes and social norms to behaviors are hypothesized to be mediated through intentions.

Several differences ought to be noted between the path model (Figure 11.6) and the latent structural equation model (Figure 11.8). First, the latent structural equation model contains a measurement component that provides information regarding the factor loadings. This information offers insights about how well each item loads on each factor and the reliability of the items that comprise that factor. Second, the factors in the latent predictive model have statistically removed the measurement error (unreliability) from each factor, whereas the path model contains factors contaminated by measurement error. Latent factors used to test predictive relations yield a more accurate assessment of the predictive relations of the constructs, as they no longer contain measurement error. Considering that measurement reliability is the upper bound of measurement validity, statistically removing measurement errors from the factors yields a more accurate view of the predictive relations.

Although a latent predictive model is considered statistically advantageous over a path model, path models are often found in the literature for several reasons. A minimum of two (preferably three) items is required to represent and assess the reliability of each latent factor. For example, if attitudes were measured with a single questionnaire item, a latent factor for that construct would not be estimable. Two or more items are needed to extract shared variance and to derive a latent factor. When fewer items are used for each measure, a path model could be estimated, with all the constructs represented by mean composites instead of latent factors. Another alternative is to estimate a hybrid structural equation model that contains some constructs represented by measured variables (such as that of a path model) and other constructs represented by latent factors (such as that of a latent predictive model).

Another reason that a path model might be reported in the published literature is that latent predictive models are more complicated to estimate and more challenging to obtain strong fit indices, which are discussed in the next section.

Fit Indices of Structural Equation Models

The statistical details to estimate structural equation models (path models and latent predictive models) are beyond the scope of this textbook and require knowledge of variance-covariance algebra or matrix algebra, but the logic of model testing can be conceptually understood as follows. The raw data are converted to a matrix of intercorrelations (or covariances) among all the measured variables used in the model. From this underlying correlation matrix, the software derives estimates of path coefficients and factor loadings based on the hypothesized model. The extent the hypothesized model corresponds with the underlying data can then be evaluated using model fit indices.

Model fit indices are generated as part of the structural equation model output and provide information regarding how well the hypothesized relations in the model correspond to the actual data as represented by the underlying correlation matrix of variables. Fit indices can be classified into two types. *Goodness-of-fit indices* typically range from 0 to 1.00, with higher values considered desirable and reflecting stronger model fit. The most reported goodness-of-fit indices are the CFI (Comparative Fit Index) and TLI (Tucker-Lewis Index). *Badness-of-fit indices* reflect the degree of discrepancy between the model and data, with low values close to 0 considered desirable in demonstrating stronger model fit. The most commonly reported badness-of-fit indices are the RMSEA (Root Mean Square Error of Approximation) and SRMR (Standardized Root Mean Square Residual). Different fit indices possess different statistical properties and different advantages and disadvantages. As such, researchers are advised to report several fit indices

to help triangulate and decide on the extent that the hypothesized model corresponds with the underlying data.

If the overall model fit indices are judged to be sufficiently strong, the researcher is then permitted to evaluate the statistical significance of the specific path coefficients in the model. If the model fit indices are poor, it is not meaningful to interpret the specific paths. In situations involving unacceptable model fit, the researcher might wish to modify the model by deleting or adding paths in an attempt to correctly specify the model and improve model fit. For example, suppose that the model in Figure 11.6 generated less-than-optimal fit indices. In the goal of obtaining better fit indices, the researcher might wish to pursue post hoc modifications by deleting nonsignificant paths and incorporating additional paths that might be significant, such as adding a direct path from norms to behaviors. According to the theory of reasoned action (Ajzen & Fishbein, 1980), it may not be theoretically justified or meaningful to add the path from norms to behaviors to the hypothesized model (Figure 11.6). The implementation of post hoc modifications must always be logically justified and transparently reported.

Even with desirable fit indices for a model, the resulting structural equation model estimates do not unequivocally “prove” that the model is the only correct one. Strong fit indices indicate that the hypothesized model is compatible with the obtained data, but other models might be even more compatible. In fact, a great number of plausible models could be generated from the same empirical data that vary in model fit indices. For this reason, it is recommended that structural equation modeling be used to test multiple competing models (i.e., different sets of hypothesized paths using the same set of variables). Fit indices cannot guarantee that one model is the only correct one in existence, but they can be used to determine which of several competing models is best supported by the data.

Modeling Longitudinal Data

Cross-Lagged Panel Models

A **cross-lagged panel model** is estimated to help determine the temporal direction of relations involving variables using nonexperimental longitudinal data. The technique is usually applied in an attempt to make causal inferences about the directional influence of processes or events. Does variable A occur before changes in variable B? Or does the opposite sequence hold? Or do both processes occur simultaneously? Cross-lagged panel analysis is considered a strategy to determine whether a variable is a temporal antecedent, consequent, or both with respect to other variables when only correlational data are available. However, it is not as methodologically rigorous as randomized experiment, and hence causal conclusions are less certain.

Figure 11.9 shows the simplest possible cross-lagged panel model. The model involves two constructs assessed across two time points, specifically to determine the temporal precedence of attitudes versus behaviors regarding caffeine consumption. A cross-lagged panel analysis requires that all the constructs of interest are assessed at every time point in a longitudinal design. In other words, to infer the prospective order of processes, each construct is estimated as both the antecedent and consequent in every measurement round. Cross-lagged panel coefficients can be estimated using structural equation modeling techniques.

Cross-lagged panel designs contain three types of statistical relations (Kenny, 1975; Lac, 2016; Locascio, 1982), as diagrammed in Figure 11.10. First, a *synchronous correlation* represents the nondirectional association of two different constructs assessed at the same time point. Specifically, T1 attitudes and T1 behaviors are allowed to correlate, and T2 attitudes and T2 behaviors are allowed to correlate. In structural equation models, it is statistically impossible to

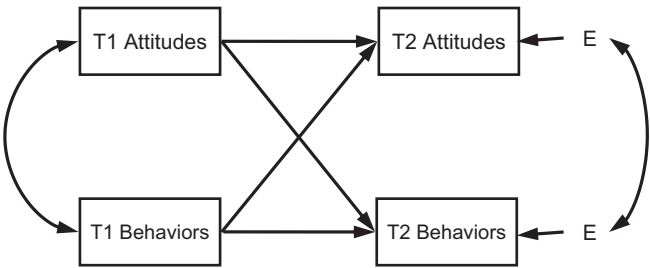


FIGURE 11.9 Cross-lagged panel model using two constructs.

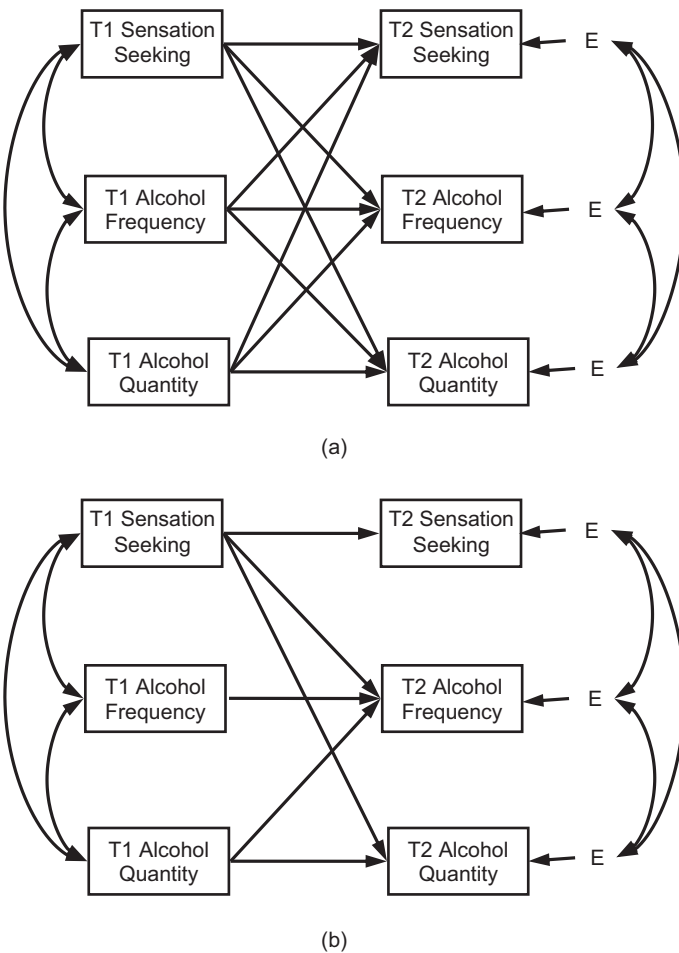


FIGURE 11.10 Cross-lagged panel model using three constructs.

correlate endogenous variables or factors. Thus, their predictive error terms are specified to correlate, as a proxy, to roughly represent the correlation between T2 attitudes and T2 behavior. Second, an *autoregression path* denotes the longitudinal association of a construct assessed in the earlier time point predicting its subsequent assessment, while statistically controlling for the other antecedent constructs in the model. The autoregression path reflects the test-retest

stability of a construct (while controlling for other constructs) and therefore inevitably emerges as the strongest coefficient in the model. The autoregression paths are from T1 attitudes to T2 attitudes and from T1 behaviors to T2 behaviors. A *cross-lagged path* of a construct in longitudinally predicting a different construct is the most pertinent for determining the temporal precedence of the variables. The cross-lagged paths are from T1 attitudes to T2 behaviors and from T1 behaviors to T2 attitudes.

The *cross-lagged paths* are interpreted to determine the temporal directionality of the relations between constructs. Several scenarios are possible. If both cross-lagged paths are found to be statistically significant, this is interpreted as a reciprocal influence in terms of mutually reinforcing processes across time. Specifically, more favorable caffeine attitudes (T1 attitudes) subsequently lead to higher caffeine usage (T2 behavior), and higher caffeine usage (T1 behavior) leads to more favorable caffeine attitudes (T1 attitudes). This pattern suggests concurrent longitudinal processes in both directions: attitudes and behavior both operate as antecedents and consequents. In a different scenario, it is possible that the path from T1 attitudes to T2 behaviors is significant, but the path from T1 behaviors to T2 attitudes is not significant. This finding would suggest a one-way directional influence such that favorable caffeine attitudes (T1 attitudes) lead to greater caffeine consumption (T2 behavior), but not the other way around. Essentially, this would indicate that attitude is an antecedent and behavior is a consequent. Yet another possible scenario is that the path from T1 behaviors to T2 attitudes is significant, but the path from T1 attitudes to T2 behaviors is not. This is interpreted as a one-way influence, such that higher caffeine usage (T1 behavior) leads to more favorable caffeine attitudes (T2), but the opposite temporal process is untenable. This finding would indicate that behavior is an antecedent and attitudes are consequent.

Cross-lagged panel models could be extended to unravel the temporal directionality across three or more variables. In a cross-lagged panel study by Lac and Donaldson (2021), sensation seeking personality, alcohol frequency (number of days of alcohol use), and alcohol quantity (number of drinks on a typical day when alcohol is used) were assessed in two time points. Sensation seeking (Zuckerman, 1979, 2007) is theorized in the literature as a personality trait responsible for the inclination to engage in novel and risky activities. Accordingly, sensation seeking might operate as a temporal antecedent, whereas alcohol quantity and alcohol frequency ought to operate as temporal consequents.

The cross-lagged panel model involving these three constructs as measured across two time points is depicted in Figure 11.10a. The three possible types of paths in a cross-lagged panel model were initially estimated: synchronous correlations (e.g., bivariate correlations involving T1 sensation seeking, T1 alcohol frequency, and T1 alcohol quantity), autoregression paths (e.g., T1 sensation seeking to T2 sensation seeking), and cross-lagged paths (e.g., T1 sensation seeking to T2 alcohol frequency; T1 sensation seeking to T2 alcohol quantity). After testing this initial model (Figure 11.10a), the nonsignificant paths were deleted to produce a final model (Figure 11.10b).

Focusing on the cross-lagged paths of the final model depicted in Figure 11.10b, T1 sensation seeking was found to be an antecedent of both T2 alcohol frequency and T2 alcohol quantity. Furthermore, T1 alcohol frequency and T1 alcohol quantity did not operate as antecedents of T2 sensation seeking. Overall, this final model offers support for the theoretical proposition of a one-way directional relation – the personality trait of sensation seeking promotes greater frequency and quantity of alcohol usage. Furthermore, the cross-lagged path from T1 alcohol quantity to T2 alcohol frequency was significant, but the pathway from T1 alcohol frequency to T2 alcohol quantity was nonsignificant. This finding shows that higher consumption in terms of alcohol quantity, as it assesses a heavier and more concentrated ingestion pattern, will foster

greater alcohol frequency in terms of drinking across days. Cross-lagged panel models could be extended to involve multiple constructs and even multiple rounds of longitudinal data.

Growth Curve Models

A **growth curve model** helps determine the longitudinal trajectory or temporal shape of a variable for participants assessed at multiple time points (Liu et al., 2012). For example, a researcher might wish to evaluate the longitudinal trend of job satisfaction for employees at a company after they were hired. Job satisfaction scores were assessed every month for 12 months. Do the results show an incremental improvement month after month, with the highest job satisfaction scores at month 12? Or do the employees only show improvement in job satisfaction from month 1 to month 6, but then the satisfaction stabilizes and scores remain flat from month 7 to month 12? Alternatively, do the employees gradually show increases in job satisfaction from month 1 to month 4, with the highest satisfaction from month 5 to month 8, but then gradually show a decline from month 8 to month 12 back to the original baseline satisfaction score? These longitudinal patterns and many other types of growth patterns may be evaluated with growth curve modeling. Although growth curves may be estimated using structural equation modeling (known as latent growth curve modeling), the focus here will be on the easier to understand multilevel modeling variant. Earlier in this chapter, the topic of multilevel modeling focused on cross-sectional data. In this section, multilevel modeling is extended to longitudinal data.

In growth curve modeling, time points of measurement are nested within each participant. The hierarchical design of growth curve modeling is depicted in Figure 11.11. At first sight, this figure may seem counter-intuitive, but it is logical once the reasoning is understood. Suppose that participants were assessed on the same measure or questionnaire every year for four years. The longitudinal points of observations (Time 1, Time 2, Time 3, Time 4) are nested within each participant. As multilevel models involve nested designs, this approach statistically controls for the clustering of temporal measurements within participants. In other words, for any single participant, that person’s nested scores across time points tend to be correlated or clustered (as the same person has completed those measures), so this ought to be statistically controlled for in the analysis. Controlling for the nested design in a multilevel model is desirable so that detection of the underlying growth trajectory is not obscured.

To illustrate some instances of growth curve models, suppose that undergraduate students at your college were longitudinally tracked over four years and their alcohol consumption was assessed every year (freshman, sophomore, junior, senior). Growth curve analysis might show that the number of drinks consumed gradually increased from year to year. This consistent incremental change from time point to time point is called a *linear trend* as depicted in Figure 11.12a.

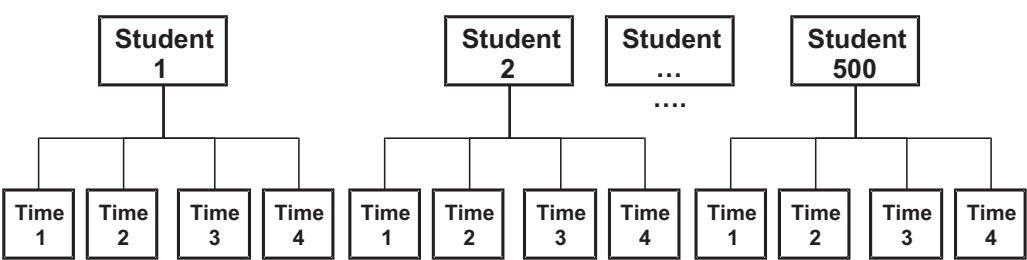


FIGURE 11.11 Growth curve model and its nested hierarchy: Measurement time points nested within each student.

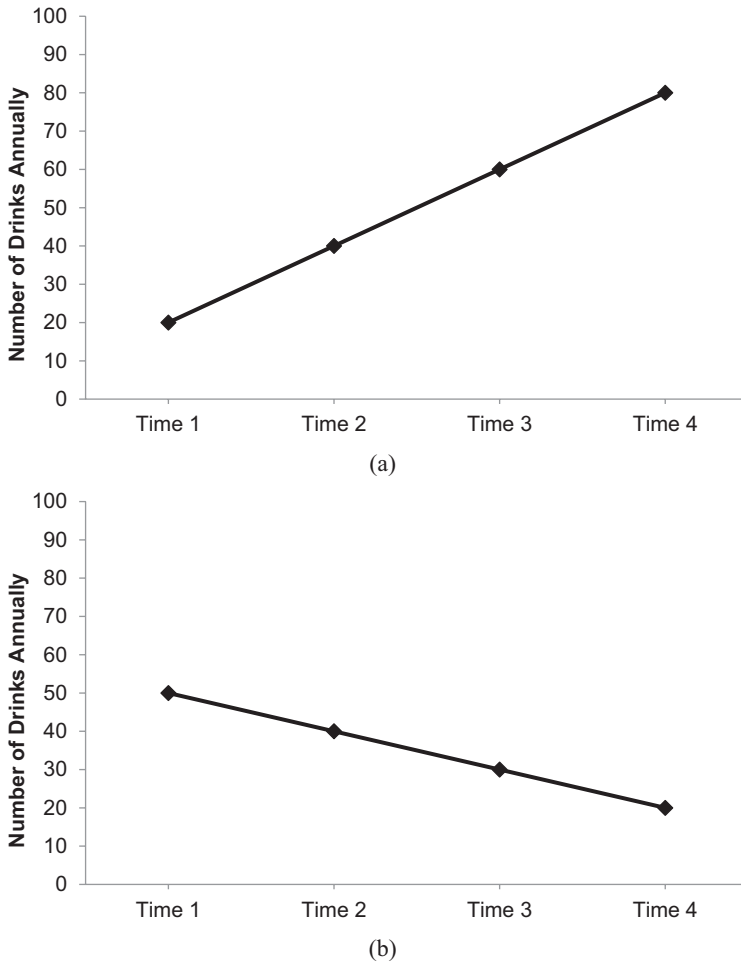


FIGURE 11.12 Growth curve models with linear trends.

Another linear trend, illustrated in Figure 11.12b, shows a consistent decrement in alcohol consumption from year to year. (This declining pattern of annual alcohol use does not reflect the reality at most colleges but may be true at some colleges.) Alternatively, it is possible for the trajectory to exhibit a *quadratic trend*, as in Figure 11.13a. Here, alcohol consumption increases from the first to second year, stabilizes in the third year, and declines to initial rates of consumption in the fourth year of college. A possible explanation for this pattern is that students start taking upper division courses during their third year, and because these courses are more challenging, they may not increase their alcohol consumption but now drink at the same rate as their second year, so they are not too inebriated to study. Afterward, the alcohol usage level in the senior year reverts down to the original rate when students reduce consumption, perhaps shifting their energy in efforts to finish the degree and graduate. Many other forms of growth trends are possible – for example, the *cubic trend* shown in Figure 11.13b.

In advanced applications of growth curve modeling, after identifying the shape of the growth curve, the researcher could incorporate predictors to help explain the trajectory of the outcome measure. For example, the predictor of gender assessed at Time 1 could be entered into the model along with alcohol attitudes assessed at each of the four time points to predict the alcohol

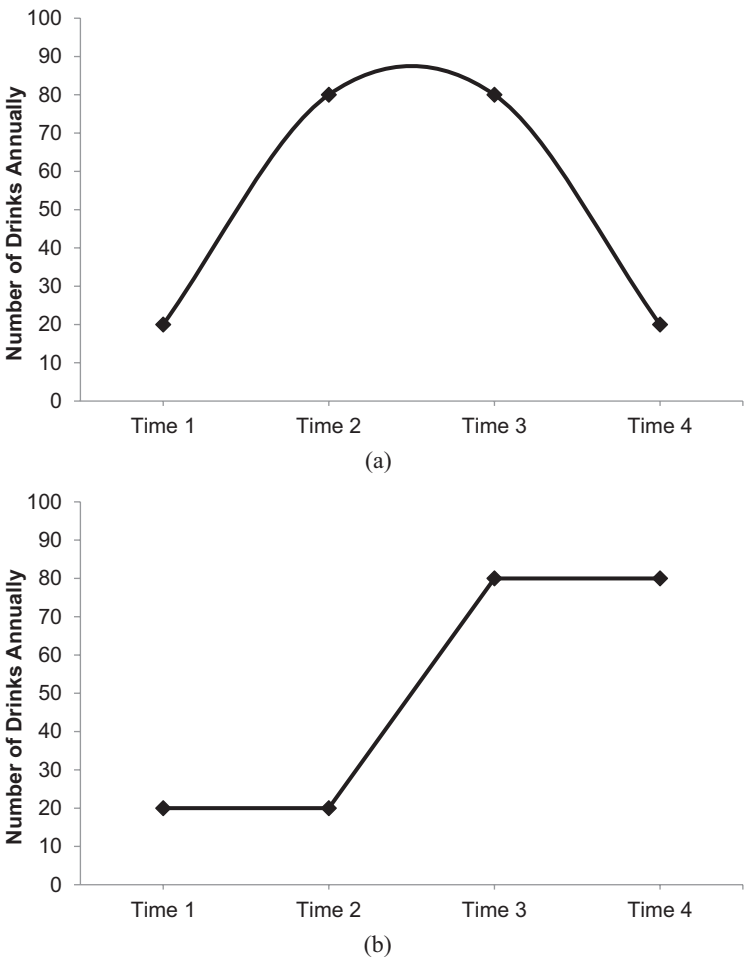


FIGURE 11.13 Growth curve models with quadratic and cubic trends.

use linear trend depicted in Figure 11.12a. Growth curve modeling permits the entry of predictors or covariates that are measured at a single time point and/or at multiple time points. Gender is considered a *non-time-varying predictor* as it was assessed at a single time point, whereas alcohol attitudes is considered a *time-varying predictor* as it was assessed across multiple time points. The results may reveal that male (compared to female) gender and increasing positive alcohol attitudes across time each uniquely predicted the linear increase in alcohol use across time. Also, growth curve modeling may be applied to compare different groups in terms of differential patterns of longitudinal change across time. Suppose the drinking patterns across four years at your college are compared to the drinking patterns across the same four years at another university that implemented a strict anti-alcohol policy. A growth curve for one group may resemble that of Figure 11.13a, and for the other group may resemble that of Figure 11.13b.

Conclusion

Although many new terms have been introduced in this discussion of multivariable techniques, much of this material builds upon the terms and concepts of multiple regression and correlation. Testing a single predictor on a single outcome, such as the case when performing a correlation,

may reflect an overly simplistic understanding of reality. Consistent with reality, many variables and processes could explain an outcome or multiple outcomes. Thus, multivariable and multivariate approaches, such as multilevel modeling, structural equation modeling, cross-lagged panel modeling, and growth curve model are necessary to estimate and understand complex interrelations of variables collected using correlational designs. The ideas that underlie correlational methods are the same as those that lie at the heart of the experimental method. In both research strategies, we begin with a theory or a set of hypotheses regarding the relations among a set of variables. These relations should be clearly stated before data are analyzed. For many, a major strength of techniques such as structural modeling is that they force us to explicitly hypothesize the relations among our variables. Having specified our model, we then collect data to test the hypothesized relations. If our theory is reasonable, the set of hypothesized relations we estimate should make sense.

One reason for caution in interpreting predictive models is implicit in our reference to the size of the predictive error terms. It is possible that a proposed model may involve variables that explain only a miniscule portion of the possible variance in an outcome variable. Such a result would be comparable to a multiple regression study in which our set of predictor variables was only weakly related to the criterion (i.e., with a relatively unsubstantial R^2). In this case, even if the result is statistically significant, it is obvious that the predictors we have chosen to measure are inadequate to provide a complete picture of the factors explicating the criterion.

The results derived from many of the multivariable and multivariate correlational approaches can never be proven correct, no matter how plausible the obtained results. Just as in the experimental methods, nonexperimental methods gain credence by a process in which alternative explanations of critical relations are shown to be less plausible than the theory under examination. Unless used in an exploratory or summative way for hypothesis development, a possibility suggested by Crano and Mendoza (1987), many of the techniques covered in this chapter involve explicit statements of hypothesized relations among variables (derived from some *a priori* empirical or conceptual framework). This is a different process from completely exploratory research, in which the patterns of associations among variables often are not hypothesized in advance but observed and interpreted after the data are collected. Interpreting results of analyses based on such post hoc observations under the guise of hypothesis-driven research represents a misuse of an otherwise powerful statistical methodology.

References

- Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Englewood Cliffs, N.J.: Prentice-Hall.
- Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173–1182.
- Berry, D. S., & McArthur, L. Z. (1986). Perceiving character in faces: The impact of age-related cranio-facial changes on social perception. *Psychological Bulletin*, 100, 3–18.
- Byrne, D. (1971). *The attraction paradigm*. Academic Press.
- Condon, J. W., & Crano, W. D. (1988). Implied evaluation and the relationship between similarity and interpersonal attraction. *Journal of Personality and Social Psychology*, 54, 789–797.
- Crano, W. D., & Mendoza, J. L. (1987). Maternal factors that influence children's positive behavior: Demonstration of a structural equation analysis of selected data from the Berkeley Growth Study. *Child Development*, 58, 38–48.
- Dion, K. K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology*, 24, 285–290.
- Heck, R. H., & Thomas, S. L. (2022). *Multilevel and longitudinal modeling with IBM SPSS* (3rd ed.). Routledge.

- Heck, R. H., Thomas, S., & Tabata, L. (2013). *Multilevel modeling of categorical outcomes using IBM SPSS*. Routledge.
- Kahn, J. H. (2011). Multilevel modeling: Overview and applications to research in counseling psychology. *Journal of Counseling Psychology*, 58, 257–271.
- Kenny, D. A. (1975). Cross-lagged panel correlation – test for spuriousness. *Psychological Bulletin*, 82, 887–903.
- Lac, A. (2016). Longitudinal designs. In R. J. R. Levesque (Ed.), *Encyclopedia of adolescence* (2nd ed., pp. 1–6). Springer International
- Lac, A., & Donaldson, C. D. (2021). Sensation seeking versus alcohol use: Evaluating temporal precedence using cross-lagged panel models. *Drug and Alcohol Dependence*, 219, 1–7.
- Liu, S., Rovine, M. J., & Molenaar, P. C. (2012). Selecting a linear mixed model for longitudinal data: Repeated measures analysis of variance, covariance pattern model, and growth curve approaches. *Psychological Methods*, 17, 15–30.
- Locascio, J. J. (1982). The cross-lagged correlation technique: Reconsideration in terms of exploratory utility, assumption specification and robustness. *Educational and Psychological Measurement*, 42, 1023–1036.
- Napolitan, D. A., & Goethals, G. R. (1979). The attribution of friendliness. *Journal of Experimental Social Psychology*, 1, 105–113.
- Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, 40, 879–891.
- Zhao, X., Lynch, J. G. & Chen, Q. (2010) Reconsidering Baron and Kenny: Myths and truths about mediation analysis, *Journal of Consumer Research*, 37, 197–206.
- Zuckerman, M. (1979). *Sensation seeking: Beyond the optimal level of arousal*. Lawrence Erlbaum Associates.
- Zuckerman, M. (2007). *Sensation seeking and risky behavior*. American Psychological Association.

PART III

Data Collecting Methods



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

12

SURVEY STUDIES

Design and Sampling

In fundamental ways, survey research is different from experimentally oriented methods that occupied our attention in the text's initial chapters. In this chapter, the term *survey* refers to the process of polling, or surveying, an identified group of respondents on topics of interest – people's attitudes, perceptions, intentions, behaviors, and so on. **Surveys** may involve mailed questionnaires, telephone interviews, face-to-face interviews, or even online pop-ups. Most survey research is intended to use obtained answers to survey items to estimate those of a larger group or population. Therefore, a central concern in surveys is concerned with the adequacy of a subset of individuals in representing those of the population from which they were drawn. Generally, survey researchers are not overly concerned with issues of internal validity (i.e., Is an experimental manipulation responsible for the obtained findings?), because surveys are primarily used in descriptive or nonexperimental research. Although some researchers use surveys as a vehicle for experimental studies, the central goal of most survey research is to provide sample estimates of population values that are as accurate as possible.

Selection versus Assignment

To draw the distinction between experimental and survey research, it is useful to emphasize the differences between respondent *sampling* (or **selection**) of participants and the *assignment* of participants to study conditions. Recall that in experimental designs, we stressed the importance of random assignment of participants to conditions. In experiments, we are concerned with ensuring participants are assigned randomly to the conditions of the study. We are less concerned with the characteristics of the original population from which participants were drawn. The core requirement for random *assignment* is that each person participating in the study has the same chance of being assigned to a specific experimental or control condition as any other person. Random assignment is essential if the full power of experimental techniques to foster causal inferences is to be realized. In *sampling*, we are concerned with how respondents got into the sample (i.e., what sampling procedures were used to identify a potential respondent as eligible to be in the study).

Most researchers recognize that the generalizability of results is dependent on the features of the larger population pool from which the participants were drawn, but this consideration is

secondary in most experimental research. For example, if participants used for a study were all students enrolled in an undergraduate course in the fall semester, then without replicating the study with demographically disparate participants, it would be dangerous to generalize findings beyond that group. Of course, we often are not content to generalize to such restricted populations, but to generalize results to other groups involves the risk of overextending the results. This should not be interpreted as a criticism of experimentation, or of random assignment of individuals selected from a restricted population. Rather, it is intended to caution researchers not to overextend the boundaries of their findings.

In experimental research, the objective is to examine different outcomes between groups as a function of an experimental treatment. To do so, participants are assigned to different study conditions. *Random assignment* of participants to different conditions is necessary to achieve high internal validity for making causal inferences. *Sampling* procedures have a different goal. They are concerned with external validity to allow generalizing results to a population of interest. There are many forms of random sampling procedures (e.g., simple random sampling, stratified, multistage) and non-random sampling approaches (e.g., convenience, snowball) discussed in this chapter, but first some important preliminary issues require consideration.

Census and Survey Basics

In the technical sampling literature, a distinction is made between a census and a sample survey. A **census** is an investigation involving *all* the potential units that could be included from the target population (the universe of interest). In comparison, a sample refers to a subset of eligible units in a population. In the United States, the government conducts a national census every decade to obtain demographic characteristics of every person living in the country. Census workers are responsible for going door to door to interview residents living in households, but to ensure comprehensive coverage of all members of the populace, homeless people and incarcerated prisoners also are surveyed.

A *census* does not necessarily imply a massive number of units, such as every person living in a particular nation (as in the U.S. census), but the intended population to which the researcher wishes to generalize the survey results. A researcher wishing to estimate the risk of contracting malaria in Kenya obviously is not concerned with malaria rates of every country of the world. Instead, the targeted population would be individuals endangered or suffering from this disease in that country. For a study examining grade point average held by teenagers living in Canada, the targeted population would be the teens living in Canada, but if the researcher's purpose is to generalize to female teenagers attending a particular high school, with findings to be used by the school administrators, then logically the population of interest is teenage female students attending that school.

In most research, a sample is preferred to the use of the complete set of possible units (i.e., a census) because, *within some reasonable degree of **sampling error***, the sample will approximate the results that would have been obtained had a complete census been taken. A sample is collected at a fraction of the cost associated with a complete enumeration of all population units (i.e., a census).¹

Precision and Sampling Error

Social researchers' preference for surveys over a complete census is based on the understanding that a survey sample will reproduce the results of a census within some reasonable degree of error. This understanding prompts consideration of an important issue in sampling, the estimate

of **precision**. An estimate is a statistical value computed from members of a sample. It serves as an inference, with some degree of sampling error, of a population value. Examples of estimates include the mathematical mean of cigarettes smoked, self-esteem score, widgets built, or the proportion of cigarette smokers, right-handers, or women in a population. The precision of an estimate is inversely related to the degree of **sampling error** (standard error of the mean), the expected *typical* discrepancy between the estimate calculated from a sample and the value that would have been obtained if the entire targeted population (census) had been used. Sampling error represents the variation in estimates expected if many samples were randomly drawn from the same targeted population. Obviously, if all the population members were included, the concept of sampling error would not be applicable. Sampling error, an estimate of the extent that a sample estimate may deviate from that of the population, should not be confused with measurement error.

Suppose we know that a particular high school contains a population of 1,500 students. The researcher is interested in the number of hours the typical student spends each academic quarter studying English composition. Figure 12.1 shows the distribution of sample means, which is the theoretical distribution of all possible randomly selected samples of a given sample size (in this case, 100), drawn from the same population. The figure shows that this population averaged 20 hours of studying per academic quarter. The distribution of sample means is found by calculating the standard error of the mean. In this case, the standard error is three hours.

To provide a concrete example of the use of sampling and sampling error, suppose we have available a list of all 45,000 students who attend a large Midwestern university. We know 20,000 of these students, our targeted population, live off-campus and we wish to determine the average textbook costs these students paid this semester. We have a list of these 20,000 students. This is known as the **sampling frame**, a listing of the population of interest from which some members of a population are drawn and used as the sample for a study. In addition to the average textbook costs, we want to know something about the distribution of costs across this population. How could we perform this study? We could perform a census by contacting all 20,000 off-campus students to learn their textbook costs in the present semester. The mathematical average of these costs across the entire targeted population could then be computed, as shown in the “Population 1” column at the bottom of Table 12.1.

Contacting all 20,000 off-campus students at our hypothetical university might tax our available research resources, so instead we decide to sample only 1,000 students from the total population. To do this, we use a table of random numbers (details of using such a table are

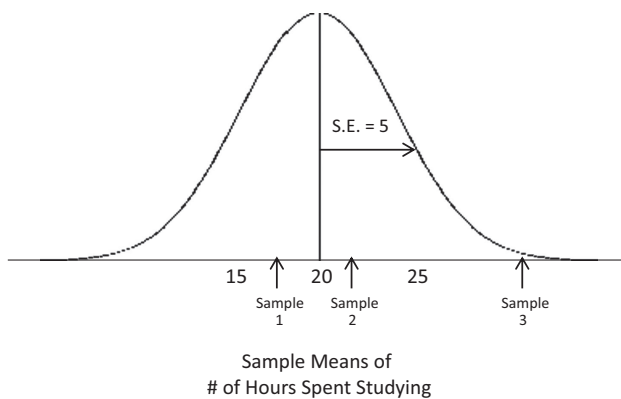


FIGURE 12.1 Distribution of sample means: 100 participants were randomly drawn for each sample.

TABLE 12.1 Textbook cost distribution of off-campus students

Textbook cost category	Population 1	Sample 1	Population 2	Sample 2
\$0–\$25	600	36	0	0
\$26–\$50	900	47	0	0
\$51–\$75	980	56	0	0
\$76–\$100	2,840	149	4,000	220
\$101–\$125	3,700	183	6,000	290
\$126–\$150	3,800	177	7,000	315
\$151–\$175	2,800	129	3,000	175
\$176–\$200	1,520	72	0	0
\$201–\$225	1,220	58	0	0
\$226–\$250	1,000	55	0	0
>\$250	640	38	0	0
Number of students	20,000	1,000	20,000	1,000
Mean cost per student	\$135.22	\$134.00	\$124.25	\$124.12
Standard deviation	57.25	59.60	24.34	25.40

discussed later) to select 1,000 people from the off-campus housing list (our sampling frame) to be part of the sample, and we contact these individuals for our survey. Usually, we will not know the true population value (called the *population parameter*); however, we will be able to estimate the parameter from the sample results. The probable accuracy between the sample and population estimate is termed the *precision* of that estimate. We can estimate the precision of a sample mean by determining the sampling error, or standard error (S.E.) of the mean. A low sampling error reflects high precision. The precision of a simple random sample is estimated by the following formula:

$$\text{S.E.} = \sqrt{\frac{s^2}{n}}$$

where S.E. = the standard error of the estimated mean (the inverse of precision),

s = the standard deviation of the sample,

n = the number of participants in the sample.²

All three terms in the formula use information from the sample to infer population value.

The formula for computing S.E. contains two clues about factors that influence the precision of an estimate. Notice that the size of the sample has much to do with the sampling error. In fact, the sample size, not the sampling fraction, plays the predominant role in determining precision. The precision formula clearly shows that as sample size increases, the standard error decreases. As a larger sample size is more reflective of the population, increasing the number of respondents will reduce the standard error of an estimate. However, the connection is not one-to-one, because it is the square root of the sample size that is entered in the formulas for the standard error. This indicates that if researchers wish to double the precision of an estimate, they will have to quadruple the sample size. Because of this law of diminishing returns for increasing sample size to enhance precision, and considering the costs of participant recruitment, most national studies (e.g., public opinion polls and national datasets) use samples of no more than 1,000 to 2,000 people to provide estimates of population values with reasonably high levels of precision.

The other important term required to calculate the formula for the standard error is the standard deviation, denoted by the term *s*, which represents the variability of the data among participants in the sample. From Table 12.1, the mean and standard deviation of the sample parallel

the corresponding values for the population. As population information usually is unknown, the sample estimates are used as an approximation of their respective population values. The larger this sample's standard deviation term, the greater the standard error of the mean. In other words, the more variable the spread of individual values used to compute the sample estimate (in our example, semester textbook costs), the greater the standard error of the sample mean, and consequently, the lower the precision of the sample mean.

Consider Table 12.1 again. As shown here, the range of costs in "Population 1" is relatively wide. A random sample of 1,000 respondents drawn from this population produced results shown in the "Sample 1" column, where the standard deviation is \$59.60. Now consider the distribution of students at another university, "Population 2." As can be seen in this example, the standard deviation of a sample (sample 2) of the respondents drawn from this population is much smaller than that of the Population 1 sample. This is so because the variability of the true population value of Population 2 is smaller than that of Population 1.

To take an extreme example, suppose that the population of off-campus students each paid exactly \$200 for their textbooks that semester. In this instance, any sample size would provide a completely accurate estimate of the population mean. A moment's reflection on the precision formulas will reveal why this is so. When values are the same for all units in a sample, there is, by definition, no variation in sample values. The sample standard deviation would equal exactly zero, and upon solving the formula, the standard error also would be zero. The results of any division of this term (no matter what the sample size, n) also would equal zero. Thus, the more restricted the sample values, the more precise the sample estimate, all other things being equal. Or, to put it another way, the smaller the standard deviation of the sample (and its population), the fewer respondents will be needed to obtain greater precision of estimate of the sample mean.

Sampling Frame

Some listing of the population must be available to obtain a representative sample of a population, and this listing is called the **sampling frame**. In simple random sampling the frame is a listing of all the individual population units (or respondents), whereas in cluster or multi-stage sampling, the initial listing involves the clusters that contain all the population units (or respondents).

Two types of sampling frames are commonly employed, *sampling frames of people* and *sampling frames of locations*. The first of these consists of a listing of all the individuals of a specified population – the students at a university, the list of registered voters of a county or precinct, all subscribers of a magazine, members of a professional organization, and so on. Such lists are convenient and can prove useful for sample selection. However, possible limitations of the list must be acknowledged. For example, suppose that we obtained a list of all *Fortune Magazine* subscribers. Would it be legitimate to use this list as the sampling frame for a survey? It would, if we were concerned with estimating characteristics, beliefs, or behaviors of this population *of subscribers*. It would be an inappropriate frame to represent the U.S. population, because the subscribers to this periodical are, on average, considerably wealthier and better educated than the typical citizen. The sampling frame needs to provide a complete coverage of the population to which inferences are to be made (see Kish, 1965).

Less obvious are the biases associated with using telephone directories as our sampling frame, a convenient, though not entirely satisfactory sampling tactic. First, Blumberg and Luke (2021) estimated that nearly 67.2% of all adults now live in homes without landline phones. And approximately 25% of them had unlisted numbers. It is dangerous to assume that a sample based on names drawn randomly from a telephone directory is representative of the general population.

Too many potential systematic differences (poverty, mobility, differential needs for privacy, latent paranoia, etc.) between those with and without listed telephones, and between listed and unlisted subscribers, make such a practice untenable. The technique of random digit dialing, discussed later in this chapter, presents one way of offsetting the problem of unlisted telephone subscribers, but it does not address the possibility that people who have telephone land lines may be systematically different from those who do not.

A sampling frame of *locations* consists of using a detailed map of a specific physical environment – a city or town, a precinct, ward, or neighborhood. Cluster or multistage sampling (to be discussed later in this chapter) are used with location sampling frames. Using maps to define the list of clusters in a sampling frame is common in surveys that seek to cover reasonably wide geographic areas. In general, however, although their use is more difficult and demanding than survey designs that employ population lists of people as the sampling frame, in many situations, no adequate population list is available. Under these circumstances, maps become a useful sampling alternative.

Random Sampling

Sampling can be classified as random or non-random (Henry, 1990). In **random sampling**, a random mechanism is used to obtain a sample intended to be representative of the underlying population. Random samples can be obtained in many ways, and may involve simple random samples, systematic samples, stratified samples, cluster samples, and multistage samples. Random sampling may involve randomly selecting members from a population, members within strata of a group, or sampling location clusters (e.g., neighborhoods in a city). Random sampling is also known as *probability sampling*, because the probabilities of a member of a population being included in the sample may be determined (at some point in the process), although probabilities of member selection may or may not be equal.

In **non-random sampling**, a non-random mechanism is used to obtain a sample from a population. Samples are gathered based on convenience, by snowball sampling, or quota sampling approaches. Also known as *nonprobability sampling*, the probability that members are selected from a population cannot be determined, usually because a sampling frame listing of potential respondents or locations is unavailable. Any approach using random sampling will produce higher external validity than non-random approaches.

Random sampling is undertaken in the service of the fundamental goals of efficiency and economy. Efficiency refers to considerations of cost and precision. One of the central preoccupations of many sampling approaches is to devise ways by which the precision of estimates can be enhanced without resorting to huge samples. Non-random sampling approaches have been developed in the service of economy and are undertaken not to enhance the precision/cost ratio, but rather to reduce the expenses involved in sampling and data collection. We will consider examples of both random and non-random sampling on the pages that follow.

Simple Random Sampling

In **simple random sampling**, every member of the population in question has an equal (and nonzero) probability of being selected every time a unit is drawn for inclusion in the sample.³ *Simple random sampling* should not be confused with the overarching term of *random sampling*. The probability of selection in simple random sampling is equal to the sampling fraction. It is calculated by dividing the number of units to be included in the sample by the total number of units in the population. Thus, in the examples of Table 12.1, the sampling fraction was 5%,

because 1,000 of a possible 20,000 students were sampled. Sampling approaches of this type are called *epsem* designs, which stands for “equal probability of selection method.” Simple random sampling, systematic sampling, and proportionate stratified sampling approaches are examples of *epsem* designs.

In creating and selecting a simple random sample, the researcher has a relatively restricted set of procedural options. In situations involving a small population, one can enter each of the population units on individual elements (slips of paper, discs, etc.), mix the elements well, and randomly select the number of elements planned for the sample. Picking names out of a hat is an example of this process: if all the names are entered individually on elements (e.g., slips of paper) of the same size, if the elements are mixed well, if the person doing the choosing does so without peeking into the hat (or otherwise exerting an influence on the particular names chosen), and if sampling without replacement occurs (i.e., the elements are not returned to the hat after being selected), we have a simple random sample. If any of these conditions are violated, the result is not a simple random sample.

In research situations in which the underlying population is large, such a process becomes unwieldy. Let’s reconsider the example of Table 12.1. Obviously, using the “name in the hat” approach would be unwise in this study. [Imagine writing the names of 20,000 people on index cards, putting these into a (very large) hat, mixing them, and choosing 1,000 cards at random.] In such cases, the use of a computer or a table of random numbers is highly recommended.⁴

The process begins with a determination of the required sample size. Guidelines for estimating sample size are presented later in this chapter. For present purposes, let us assume as in the example that we have decided upon a sample of 1,000 students from the total eligible population of 20,000. To select the specific students who are to constitute our sample, we would number each of the names on our population list, from 00001 to 20,000. This list is our sampling frame. Then, using a table of random numbers, we could select the first 1,000 different five-digit random numbers that corresponded with the numbers on the list of students. So, if we came upon the number 18,921 when searching in our random number table, we would include the 18,921st student on the list in our sample; however, if a random number of 22,430 was obtained, it would not be used to select a unit into the sample, because there are only 20,000 eligible students on our frame (i.e., there is no student whose “ID number” corresponds to that from the random number table). When employing a published random number table, it is good practice to pick a starting point at random each time the table is used. This helps to assure the same investigator does not always make use of the same set of random numbers when selecting samples.

A list of random numbers may be generated on www.randomizer.org. Many potentially useful approaches for drawing simple random samples have been suggested for reasons of ease in various sampling contexts, but if our recommendation were sought about the appropriate way of drawing a sample, it would be very simple: use a random number application.

Systematic Sampling

An alternate means of choosing the students from our off-campus students list involves **systematic sampling**, which involves sampling every predetermined *n*th member from a population. In this approach, as before, a specific sample size is determined. Then, the size of the sample is divided by the total eligible population to determine the sampling fraction. In our example, the sampling fraction was 1,000/20,000, or 1 in 20. A number between 1 and 20 is obtained from your random number generator, and then every 20th person after that number in your sampling frame is selected until 1,000 respondents are chosen. Thus, if we randomly chose the number 15 as our starting point, we would include in our sample the 15th, 35th, 55th, 75th,

etc., student from the textbook cost example. We would continue in this fashion until we had sampled exactly 1,000 students.

Systematic sampling resembles simple random sampling, because all the units in the sampling frame *initially* have an equal chance of being selected (systematic sampling is an epsem method). It differs from simple random sampling because the probability of units being included in the sample is not equal. Thus, if the number 15 were randomly chosen as our starting point, the probability of the 16th student being included in the sample is zero, because our sampling interval is every 20 units. However, the probability of students 15 and 35 both being included in the sample is $1/20$, because if 15 is chosen (a 1 in 20 chance), then 35 is sure to be chosen as well. This technique is classified as a type of random sampling because the sampling begins with a random starting point.

In practice, it is generally assumed that if the listing of population members in the sampling frame is haphazard, or unsystematic, the resulting (systematic) sample approximates a simple random sample. This generally is a safe assumption unless the members of the frame are ordered in a systematic or cyclical fashion and the sampling interval coincides with the length of the cycle. For example, a frame contains the names of all heterosexual applicants for a marriage license, and the names of the couples to be married are listed in order, with the woman's name always coming before the man's. If we used an odd number as our sampling interval, our sample would be evenly represented by women and men. However, an even numbered interval would produce a sample consisting entirely of women, or of men depending on the starting point. This kind of regularity precludes obtaining a representative sample. However, sampling frames with cyclical arrangements are rarely met in practice, and situations in which they may occur are usually easily recognized (Kalton, 1983). Systematic sampling was used in earlier times because it was more convenient than using a table of random numbers. That is no longer so, given ready availability of true random number generators.

Stratification and Stratified Samples

Moser and Kalton (1972) made the important point that the definition of randomness refers to the way a sample is drawn, not to the outcome of the process that yields the sample. Thus, on rare occasions, it is conceivable that one could draw a random sample that, in fact, appears anything but random. To return to our textbook cost example, suppose that one of our research issues concerned the question of whether student costs were associated with class standing. Do off-campus seniors, for example, typically pay more for books than off-campus sophomores? To answer this question, we would require that both seniors and sophomores be included in the sample. However, it is possible that even if there were a fair number of sophomores on our list, our sample might contain none. Such a chance event of highly imbalanced subsamples is extremely unlikely in a large sample, but if we were extraordinarily unlucky, it could occur, even if our sampling technique were flawless. Small sample sizes are considerably more susceptible to this problem.

To ensure that an adequate number of participants are selected from each of the different subgroups of the population, survey researchers have avoided this issue by using a technique known as **stratified sampling**, in which the population is divided into theoretically meaningful or empirically important strata before members are randomly drawn from each stratum (or subpopulation) and used for the sample. Respondents then are randomly selected *from within each stratum*, and this permits prespecified subsample sizes for each stratum. The subsamples are later combined to yield the final sample. Two forms of stratified sampling are common, depending on the way the sampling fraction is employed. Frequently, the same sampling fraction is used

for each of the strata; in such a case, the result is called a *proportionate stratified (random) sample*. Sometimes, a different sampling fraction is employed within each stratum; in this instance, the resulting sample is termed a *disproportionate stratified (random) sample*. Generally, research issues dictate whether proportionate or disproportionate stratified sampling will be used.

To provide an example of contexts that would call for the use of proportionate or disproportionate stratified sampling, consider the following scenario. Suppose the Democratic governor wanted to raise the state income tax but was concerned about the effect such an action might have on her political future. To obtain prior information regarding the effect of such a move on her popularity, and the differential impact that such a proposal might have on a number of different “constituencies” that were important to her continued effectiveness, the governor commissions a survey to study the issue. Depending on the number and complexity of the populations (constituencies) of interest, this survey could take many forms, but let’s begin with a simple case.

Suppose we wanted to determine the reaction to the tax hike of people from two obvious constituencies, registered Democrats and Republicans. To survey and ensure adequate coverage of both subpopulations, (1) the list of registered voters in the state (the sampling frame) could be divided into Democratic and Republican strata, (2) some sampling fraction (say, $1/1,000$, or $.001$) decided upon, (3) the appropriate number of people for the sample of the total number of registered Democratic and Republican voters randomly chosen from the combined lists, and (4) the resulting sample polled. Based on the 2020 election, the registered sample would yield 947 Democrats and 836 Republicans. Because the same sampling fraction is employed for both strata, the sample is a proportionate stratified *random sample* (i.e., an “epsem” design). Note that in this approach, the proportion of Democrats and Republicans is the same in the population and the sample (the sampling fraction is $.001$). Thus, Democratic voters constitute 53% of the population, and so 53% of the sample would consist of registered Democratic voters. Proportionate stratified random sampling can increase precision if the stratification variable makes a difference. If Republicans and Democrats react differently to the tax increase idea, their conflicting responses would be treated as error variance had political affiliation not been included as a stratifying variable. Stratifying helps account for some of the systematic differences associated with party affiliation.

To embellish our example, suppose the simple Democratic/Republican categorization was too gross for the governor’s purposes. A more fine-grained voter stratification was desired. To do this, Democrat and Republican voters could be further subdivided by biological sex of respondent. The resulting sample, formed by the combination of the selection variables – political party affiliation and biological sex – using a constant or uniform sampling fraction (again, let us say, $.001$), could be surveyed and their data analyzed. The findings would supply a more precise overall estimate of the state’s electorate than the simpler approach, but it would be considerably more costly. Given the use of the same sampling fraction or proportion from each subpopulation, the sample takes the form of a proportionate stratified sample.

Oversampling. In some circumstances, proportionate stratified sampling does not provide enough respondents within a stratum to allow for confident statistical analysis. For instance, a survey using this sampling approach could result in relatively few respondents in a subsample. To offset this problem, the smaller group might be studied by using a sampling fraction greater than one that could otherwise be used in a proportionate stratified. A larger sampling fraction would be used to assure a number of those individuals from the small group sufficient to allow for a sample large enough to provide a good estimate of the subpopulation. This approach is said to use a disproportionate stratified sample, and as noted, the reason for oversampling is to provide enough respondents on which to base a surer estimate of a population value.⁵ This approach may

be pursued if participants were sampled in terms of their gender: males, females, and nonbinary. As people who identify as nonbinary are relatively uncommon in the population, a researcher may wish to oversample the stratum of nonbinary participants so that they will be adequately represented in the sample.

Why stratify? Why should the survey researcher bother to stratify? One compelling reason is to offset the selection of unusual or nonrepresentative group of respondents from a subpopulation. Although this non-representativeness is always possible in survey research, it is unlikely if members of a subpopulation are reasonably well represented in the population. More importantly, proportionate stratification helps ensure the distribution of respondents of each group in the sample is approximately the same as that in the population, and this enhances the precision of our estimate. In other words, by adequately representing the subpopulations in our sample, we reduce sampling error in our estimates, because formerly unaccounted variance is now accounted for by the stratification variable – party affiliation, gender, or other variables that might be theoretically or practically relevant, such as socioeconomic status, religion, members of state militias, etc.

Just as stratification by political party differences enhances the precision of estimates, so too does stratification on other factors that might be used. In general, the more stratification factors, the greater the precision of estimates. Stratification enhances precision to the extent that respondents of different strata characteristically respond differently on the issue under investigation. Thus, returning to our voting example, if the sample of Democrats and Republicans responded identically to a tax increase proposal, then stratifying on party affiliation would have no effect on precision – the stratification factor would not remove any source of variance. If they responded very differently, however, then stratification would positively affect precision of estimates. The trade-off bears consideration – stratification can be expensive and difficult, so to determine whether it is worth the costs involved, we should first decide whether the stratification factor is in some way related *systematically* to the variables under study.

Cluster Sampling

In many research contexts, the available sampling frame does not provide a list of all population members, but rather a location frame consisting of *places* in which population members might be found. For example, suppose we wished to sample the parents of public high school students in Brooklyn, NY. We are interested in their feelings about the availability of advanced placement courses. The school board is reluctant to give us the list of names of the parents/guardians. Thus, we have no sampling frame, at least initially. However, we could proceed by obtaining a detailed map of the city, which listed each block, and then randomly sampling these natural clusters to use as the sampling frame. In **cluster sampling**, geographic locations (clusters, or segments) are randomly sampled, and all members from the selected clusters are used for the sample. In this method, the sampling frame is identified (say, all city blocks of houses), and from this population, specific clusters (city blocks, in this case) are chosen randomly. Once a cluster is chosen for inclusion in the sample, *all* members of the cluster are surveyed (in our example, all eligible parents of high schoolers within the chosen cluster, or block, would be surveyed). Cluster sampling is classified under the umbrella of random sampling because the clusters are randomly chosen, although all members within the selected clusters are then used for the sample.

Multistage Sampling

In **multistage sampling**, clusters of locations are randomly sampled from a geographical sampling frame (as in cluster sampling), and then (unlike cluster sampling) participant units within

each cluster are randomly sampled as well. As the name of this approach suggests, the sampling process is extended to more than one stage or occasion. To return to our high school example, for a two-stage sample, we divide the total city into blocks, and then randomly select some of these blocks for our sample. However, instead of including all members of the chosen blocks (i.e., cluster sampling), we randomly sample from within them. Thus, only members selected from each of the chosen clusters are used. This procedure can be extended to more stages. For example, in a national election survey, we might break the nation into counties, then census tracts, then blocks within the chosen census tracts, and, finally, households within the tracts, and then specific voters within the household (the oldest member of the household, or the member over 21 years of age whose birthday is next). Figure 12.2 illustrates the differences between cluster and multistage sampling. As presented in the figure, random sampling takes place exactly once in cluster sampling, with clusters being randomly selected, and all individuals within these selected locations are included. Multistage sampling involves two stages, and random sampling takes place more than once, at the cluster level and again at the participant level of the selected clusters.

A potential distorting influence in cluster or multistage sampling is that clusters generally do not contain the same number of potential respondents. For instance, in our high school example, not all the selected city blocks contain the same number of respondents. This problem is exacerbated when the cluster sizes are of great variability, as would be the case when counties are used as the primary sampling units in a state-wide or national sample, and individual households the elements to be sampled. In such cases, potential respondents do not all have an equal probability of selection, and the precision of the resulting estimates is thereby jeopardized.

An approach known as *probability proportional to size sampling* (PPS) has been developed to solve this potential problem. Kalton (2020; also Levy & Lemeshow, 2008) has discussed the details of this technique; for our purposes, it is sufficient to understand that the PPS sampling approach ensures that the likelihood of selection in a cluster or multistage sample is the same for all potential sampling units (or respondents) no matter the size of the cluster from which they

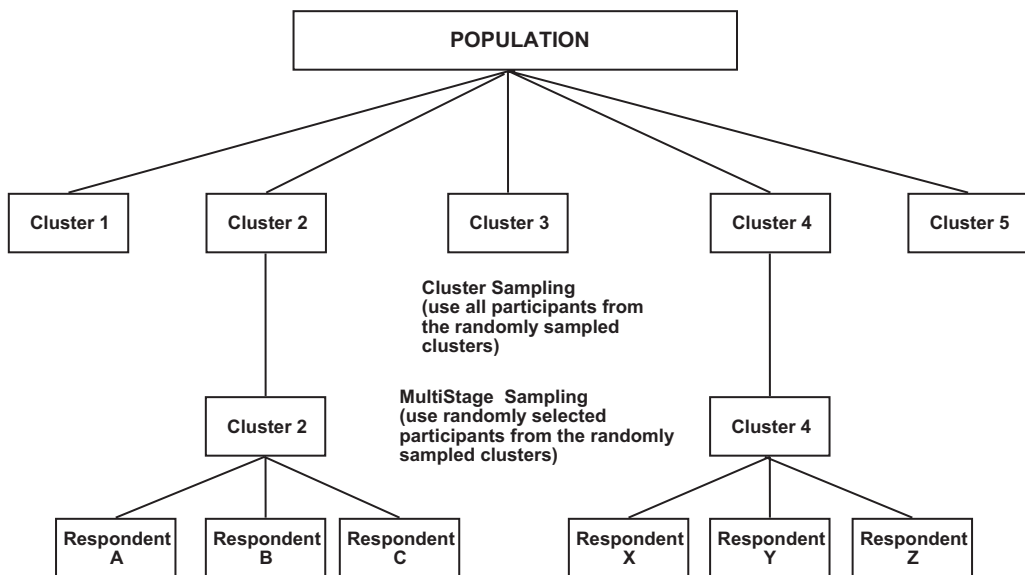


FIGURE 12.2 Schematic diagram of cluster and multistage sampling.

are drawn. Under such constraints, an epsem sample emerges, and the standard approaches for estimating precision may be used. Generally, PPS sampling is preferred in cluster or multistage sampling designs.

Multistage sampling is particularly useful when the population to be studied is spread over a large geographic area. Suppose we wished to use face-to-face interviews to determine the proportion of a state's inhabitants who used some form of state-financed social service. Using a multistage sampling approach, we could use the state's natural segmentation into counties, and randomly select a given number of counties for the sample. Then, we could randomly select towns within each of the chosen counties, neighborhoods within the towns, and blocks within the neighborhoods. Then, we could list the dwellings within the blocks, and (in the case of multistage sampling) sample individual respondents within the dwellings.

Notice that by using this approach, there is no necessity for constructing a sampling frame of individuals, which in many instances would be prohibitively expensive, if not impossible. Instead, our interviewers need only learn the number of potential respondents within a household selected for sampling, and then select among respondents by some prespecified scheme. Some survey samplers, for example, use the "birthday" rule: The person whose birthday is closest to the date of the interview is chosen as the respondent. In research contexts of this type, multistage sampling offers a practical alternative to simple (or stratified) random sampling, which requires a sampling frame of potential respondents before the initiation of the research. However, like cluster sampling, it does require a sampling frame of locations. The multistage sampling approach usually does not provide estimates that are as precise, and this is especially true if the clusters are homogeneous on the issues tapped in the survey.

With cluster and multistage sampling approaches, the precision of the survey estimates depends on the distributional characteristics of the variables of interest. If the population clusters are relatively homogeneous on the issues that are central to the survey, with high heterogeneity between clusters, the results obtained through this method will be less precise than those obtained from a random sample of the same size. However, if the population clusters are relatively heterogeneous (i.e., if the individual clusters provide a representative picture of the overall population), multistage sampling can generate estimates as precise as simple random sampling of respondents. Notice that the heterogeneity of a sample has different implications for cluster and stratified sampling. With stratified sampling, increased heterogeneity within strata results in a *loss* of precision – the respondents (or sampling units) within strata ideally should be as similar as possible. With cluster or multistage sampling, however, increased heterogeneity within clusters results in *increased* precision.

Two-Phase Sampling

In two-phase (or double) sampling, all respondents (who might have been chosen by any sampling method, such as simple random, cluster, or multistage sampling) complete the basic survey. Then, either concurrently or some time thereafter, additional information is sought from a subset of the previously selected respondents from the first stage.

Two-phase sampling is useful and widely used. The United States Bureau of the Census, for example, has used two-phase sampling in its population counts for more than 40 years. In the census, standard demographic data are collected from all households, but some households are randomly selected to be asked to provide considerably more information in the form of a more extensive survey.

Perhaps the most important use of two-phase sampling is in developing stratification factors. When the auxiliary data are collected after (rather than concurrent with) the basic survey

information, the initial survey can be used to create stratification factors. Often in two-phase sampling, numerically rare groups that cannot be identified based on readily available information are sought for study. Based on the information obtained in the first-phase sample, the sought-after group is identified, and usually disproportionately oversampled in the subsequent phase, with a lesser sampling fraction being used for the remaining population. In this way, rare respondent groups can be identified in sufficient numbers to allow for their study.⁶

Panel Surveys

The two-phase sampling method must be distinguished from the **panel survey**, in which a pre-specified sample of respondents (a panel) is surveyed repeatedly, over time. In panel surveys, the respondent sample is not successively “whittled down” over the course of the sampling phases, as is characteristic of multiphase sampling. The purpose of the panel survey is to assess individual changes that might occur in the knowledge, opinions, or perceptions of the respondent sample over the course of time. Panel surveys have proved especially useful in public opinion contexts, in which the performance of a political figure is monitored over time, in surveys of voter preference (e.g., Rahn et al., 1994), adolescent drug use (e.g., Crano et al., 2008), and in studies of different ways of boosting response rates in panel surveys (Patrick et al., 2021). Blalock (1985) provides a useful discussion of some of the applications and methodological difficulties that panel designs can involve. Given careful sampling, panel surveys can provide a sensitive index of the mood of the entire nation. Indeed, panel surveys have proved to be very powerful tools in predicting the shifting allegiances of the electorate, and the consequences of such longitudinal shifts on the outcomes of state and national elections.

Non-random Sampling

The probability sampling approaches serve the primary goal of obtaining a sample intended to be optimally representative of a targeted population by using some form of random selection mechanism (Henry, 1990). Extending results from a specific sample to a broader population is not the primary concern when using non-random sampling. Usually in non-random sampling, respondents are obtained in any way possible, so that an adequate sample size and sufficient statistical power is achieved. Unlike random sampling strategies, sampling frames are unavailable or unobtainable by the researcher. Although generalizability of results to the larger targeted population is uncertain with non-random approaches, data derived from such samples are more informative than had no information been collected at all. As part of the larger scientific enterprise, the hope is that the study results would be corroborated subsequently by other researchers, through cross-validation or replications using different (and perhaps better) samples with varying characteristics. Non-random samples are found frequently not only in the social sciences but in medical sciences as well. For instance, in biomedical research, a new antidepressant drug may be tested on a volunteer sample consisting of people aged 60 or more. Based on this sample, the results might show efficacy, which meets FDA approval, and it is released into the market. The volunteer sample is not reflective of the typical person, health status, or demographic characteristics of the larger population of people suffering from depression. However, delaying the release of this drug to others who may benefit for the sake of external validity would be considered unwise. After the drug has been on the market, other researchers might come along to evaluate whether its efficacy holds in other samples, such as young men and those with high cholesterol. Clearly, non-random sampling methods are limited in external validity. However, they are widely used, particularly for nonexperimental research where correlations between variables are

of interest. Let us consider three commonly used approaches: convenience sampling, snowball sampling, and quota sampling.

Convenience Sampling

In **convenience sampling**, the sample consists of people most readily accessible or willing to be part of the study. Volunteers or students in introductory psychology courses are commonly employed in convenience samples. Convenience samples using undergraduate students are especially common in the social sciences. Undergraduate students are not representative of the population at large. They are not even representative of people their age, as those who attend college are systematically different (e.g., higher IQ and SAT scores, higher socioeconomic status, heavier use of alcohol, etc.) from those who do not. Furthermore, the use of undergraduates at a particular school is unlikely to be representative of all college students in general. College student samples might not even be reflective of the student body, as first- and second-year students may be overrepresented in introductory courses from which such samples are derived.

Despite the potential of the Internet as a tool to sample individuals from all walks of life, most research collected through this format also makes use of convenience samples. Even if everyone who visits a website is invited to participate in a brief survey, those who click on a link to accede to the request for participation are more prone to be regular visitors of that website, have an interest in the survey topic, and usually have more free time to participate.

Using a convenience sample, Hazan and Shaver (1987) examined interpersonal relationship styles with a quiz posted in a local newspaper, the *Rocky Mountain News*. The questions were printed in the Lifestyles section with a captivating headline, “Tell us about the love of your life.” Readers were asked to choose which of three categorical descriptions best described their feelings about romantic love. The 620 respondents who mailed in responses were subsequently classified into one of the three categories of romantic attachment styles: 56% secure (i.e., comfortable with intimacy), 25% avoidant (e.g., disliked intimacy), and 19% anxious/ambivalent (worried about intimacy). The results of this convenience sample are unlikely to reflect the population of all people who have had a romantic relationship. It probably was not representative of newspaper readers in general, or even the readership of this local paper. As the headline was posted in the Lifestyles section, it was not startling to find that more than two-thirds of the respondents were female. Convenience samples are suggestive, but far from definitive. Even so, since Hazan and Shaver’s seminal study, other investigations examining the distribution of attachment styles, using different sources of recruitment and demographically different samples, have supported the findings from the original research. In fact, using a nationally representative sample, Mickelson et al. (1997) closely approximated the original results and found that the most common romantic attachment style was secure (59%), followed by avoidant (25%), and then anxious (11%).

Snowball Sampling

In **snowball sampling**, initially sampled participants are asked to contact and recruit others in their social network. These new recruits in turn are asked to invite additional members, and so forth. Those who know initially recruited individuals will have a greater likelihood of being included in the sample. Snowball sampling is particularly useful as a technique to recruit individuals belonging to marginalized or hard to obtain groups. For example, one investigation on post-traumatic stress disorder obtained a sizable sample of defectors from North Korea by first contacting South Korean welfare and vocational agencies and support groups (Chung & Seo,

2007). These initially recruited defectors were asked to invite other defectors they personally knew to be involved in the research, and then this new set of participants were asked to recruit other defectors, and so forth. Snowball sampling also has been profitably used to investigate the connection between childhood traumas and suicide attempts using a beginning sample of homeless street youth who were then asked to help recruit additional children they knew from the streets to be get involved in the study, and so forth (Hadland et al., 2012).

It is sometimes necessary to offer financial incentives to induce respondents to dig through their personal contacts and invite friends and families to serve in a research project. A sample furnished through snowball sampling tends to yield a final sample of respondents who are more like each other than would be expected by chance. Respondents recruited through snowball sampling will tend to have acquaintances who tend to be personally alike (e.g., politically, racially, living in the same city) compared to typical members of the intended larger population.

Quota Sampling

In **quota sampling**, members are sampled non-randomly until a predefined number of respondents for each of a specified set of subgroups is achieved. It differs from stratified sampling, which requires a listing of all members from each subpopulation, which is used to select members from each subsample. As a list of all subpopulation members of the targeted population is not always obtainable, the researcher might resort to quota sampling.

In some types of quota sampling, quotas are set relative to a rough guess of the subgroup proportion in the population. In other types of quota sampling, a number is set to achieve sufficient statistical power for each of the subgroups. Using this latter approach, for example, a researcher using quota sampling might wish to test the average time it takes men and women to solve a Sudoku puzzle, with the targeted population being math professors. To recruit respondents, the investigator stands outside the math department at several universities and solicits volunteers. Quota sampling is used as math professors are predominantly males. After a day of data collection, the researcher may have administered the puzzle task to 50 male math professors but only ten female math professors. Thus, sampling comes to a halt for men as the quota has been satisfied but continues for women until the number in that subgroup hits 50. In this scenario, the investigator is not concerned with female and male math professors proportional to the gender distribution of math professors in the targeted population, but the difference in mean scores in terms of minutes to solve the puzzle between male and female math professionals. By setting a quota to ensure an adequate number of members of a limited or rare subgroup are included in the sample, the calculation of the estimate for the rare subgroup will be more stable.

Other Sampling Issues

Sample Size

The question of sample size must be addressed when developing a survey. How many respondents must we sample to attain reasonable precision when estimating population values? Before we can answer this question, we must answer a series of other questions, and once these issues are decided upon, the sample size can be defined. As in all other areas of survey sampling, the decision regarding the size of the sample involves trade-offs, most of which concern the complementary issues of cost and precision.

It seems intuitively plausible that the size of the sampling fraction would be an important determinant of the adequacy of the sample. Such is not the case, however. When the underlying

population is large, precise sample results can be obtained even when the sampling fraction is quite small. What matters most is the *absolute size* of the sample, rather than the size of the sample relative to the size of the population. A moment's reflection on the formula for the standard error of a simple random sample will show why this is so. This formula demonstrates that the size of the sample, not the sampling fraction, determines precision.

The first decision that must be made in determining the survey sample size concerns the amount of error we are willing to tolerate. In general, the greater the desired precision, the larger the sample needed. Thus, if we wish to obtain extremely precise findings – results that will estimate underlying population values with a high degree of accuracy – it must be understood that we will need to sample more respondents. Cost and precision go hand in hand. Suppose we wish to estimate the proportion of the population of those who possess a specific trait or characteristic, or hold a given belief or value. Having made the decision regarding the degree of precision necessary for our purposes, we must make a rough estimate of the proportion of the population that possess the trait or belief. For example, we might want to estimate the proportion of a city's electorate that would vote for the Democratic gubernatorial candidate. To estimate the necessary sample size with a defined degree of precision, we would first have to estimate this percentage. Suppose that prior elections or recent informal surveys indicated that Democratic candidates could be expected to garner approximately 58% of the vote in the city. Having decided upon the necessary precision, we could determine the necessary sample size for a simple random sample by the formula,

$$N = \frac{p(1-p)}{(S.E.)^2}$$

where N = the necessary sample size,

p = the estimated proportion of the city's population who plan to vote Democratic, and

$S.E.$ = the sampling or standard error of the sample proportion (i.e., the amount of error we can tolerate).

Applying this formula, assuming a standard error of 5% as acceptable would yield the following:

$$N = \frac{.58(1-.58)}{.05^2} = 97$$

Thus, 97+ respondents are necessary. It is unlikely in most contexts that a standard error of this magnitude would be acceptable as it results in a 95% confidence interval with a margin of error of $\pm 5\%$.⁷ This indicates that we can be reasonably sure the actual electoral results would fall within (plus or minus) 5 percentage points of our estimate.

This example provides a useful point of contrast for later calculations. It is informative to manipulate the two crucial components in the equation to determine the effects of differing precision needs, and the influence of the accuracy of the proportion in estimating the population value on sample size. Suppose that instead of a standard error of 5%, the governor decided that no more than a 2% standard error was acceptable. This is a more reasonable choice, as it results in a 95% confidence interval with a margin of error of $\pm 2\%$. In this case, the size of the simple random sample necessary for greater precision in the sample estimate reflecting the population value would change from 97+ people to 610. That is, 610 respondents would be needed to arrive at an estimate with a 2% standard error. If an even smaller standard error of 1% were sought, as might be desired in a close race, more than 2,400 respondents would have to be surveyed. Again, the interplay of recruitment cost and precision of estimate is obvious.

A word of caution is required at this point. The calculations of the size of a sample necessary to estimate a population parameter at a given level of precision will hold only if the sampling procedure itself is properly executed. If our basic sampling technique is flawed, then the number of respondents sampled will not matter much – the results of the sample will not accurately reflect population values. A classic example of this observation is provided in a real-life case history that many think set back the cause of survey research for many years. This case, an often-told example in elementary sampling courses, occurred in 1936. At that time, a popular magazine, the *Literary Digest*, attempted a national poll to estimate who the electorate would choose as its next president, Franklin D. Roosevelt, the Democratic candidate running for his second term, or Alf Landon, the Republican nominee.

The results of the poll, printed in bold headlines at the beginning of the October 31, 1936, edition of the *Digest*, read:

Landon, 1,293,669; Roosevelt, 972,897
Final Returns in The *Digest's* Poll of Ten Million Voters

Obviously, the sample size used by the *Digest* was more than respectable -- it consisted of more than ten million people! Yet, the *Digest's* estimate was not only wrong, it was laughably misguided. Despite the *Digest's* prediction, Roosevelt ended up with 523 electoral votes to Landon's 3! As a result, the headlines of the November 14 *Digest* read,

What Went Wrong with the Polls?
None of the Straw Votes Got Exactly the Right Answer – Why?

What did go wrong? Obviously, we cannot fault the *Digest* based on sample size. But just as obviously, the poll's results were wildly off base. A brief consideration of the sampling procedures used by the *Digest* supplies insight into these problems, and illustrates the fact that sample size is not sufficient to guarantee accuracy of population estimates.

To perform their survey, the *Literary Digest* used telephone books and its membership list as the sampling frame. From their list, the *Digest* randomly sampled a large number of potential respondents and sent them a “ballot” that was to be returned by mail. In retrospect, there are two glaring problems with this procedure: first, in 1936, the proportion of the populace that owned telephones was not nearly as great as it is today. Those who did own phones were among the wealthy, a group that traditionally votes Republican. What's more, only 20% of the mailed ballots were returned. Again, there is evidence that suggests wealthier, better educated people would do so. (In fact, the *Digest's* own past polling experience had demonstrated that this was the case.) As such, the *Digest's* sample was fatally flawed. It grossly over-sampled those who, based on important demographic indicators, would be likely to vote for a Republican presidential candidate, and grossly under-sampled those who would likely vote Democratic. Under these circumstances, it is little wonder that the *Digest* predicted the Republican candidate would receive a landslide 57% of the vote when, in fact, he received only a fraction of this proportion. The kinds of errors made by the *Literary Digest* are the stuff of which sampling legends are made.

Nonresponse

Attributable at least in part to the *Literary Digest* fiasco is the problem of drop-outs or nonrespondents. Because sampling theory is based on probability theory, the mathematics that underlie sampling inference assume *perfect* response rates, that those identified for selection into

the sample will fully respond to the survey. Otherwise, sampling weights must be derived and applied to the estimates. As Moser and Kalton (1972, p. 166) observed,

The theory is based essentially on the textbook situation of ‘urns and black and white balls’, and, while in agricultural and industrial sampling the practical situation corresponds closely to its theoretical model, the social scientist is less fortunate. He has to sample from an urn in which some of the balls properly belonging to it happen not to be present at the time of selection, while others obstinately refuse to be taken from it.

Survey researchers have devised ingenious methods to reduce nonresponse in data collection. In mail surveys, for example, cash incentives to participate and follow-ups sent to those who do not return the survey within a specified period have been shown to have a positive effect on the response rate (e.g., Church, 1993; Griffin et al., 2011; Trussel & Lavrakas, 2004). More recent research has focused on enhancing executives’ response rates (Cycyota & Harrison, 2002), recruiting women veterans into suicide prevention programs (Gaeddert et al., 2020), and using incentive and reminder strategies in research involving physicians (Cook et al., 2016). In addition, considerable care is exercised in developing questionnaires of reasonable length, in personalizing the survey for the respondent, and in guaranteeing anonymity and confidentiality. In general, these tactics seem to be effective in raising response rates.

In telephone and face-to-face interviews, we encounter problems somewhat different from those of the mail survey. In these interview surveys, the researcher must distinguish “not-at-homes” (NAHs) from refusals. In the case of NAHs, researchers often repeatedly attempt at different times of day to contact the potential respondent. With refusals, different tactics are used. More established survey research organizations employ people with the descriptive job title of “refusal converters,” whose job is to persuade recalcitrant respondents. In general, in the typical phone survey or face-to-face interview, most attempts to reduce the refusal rate focus on the means to gain “entry” or to develop rapport (see Chapter 15 on interviewing for an extended discussion of these issues). Many phone surveys, for example, attempt to induce respondents’ compliance by presenting a set of nonthreatening, impersonal, and innocuous questions at the beginning of the survey. Having obtained the respondent’s initial buy-in, the interviewer then moves on to less innocuous disclosures, which people often are reluctant to provide at the beginning of the interview.

When the chosen methods to secure cooperation fail, as they sometimes do, survey analysts move to a different strategy, which involves the attempt to reduce the biasing impact of those who have been selected for the sample, but who are not present when the interviewer tries to contact them, or who refuse to cooperate. Detailing the technical aspects of these approaches is beyond the mission of this chapter; however, some of the more common techniques conceptualize the response/nonresponse distinction as a category on which members of the population can be stratified. If the researcher can obtain some information about the nonrespondents and compare it with that of the “respondents” stratum, then estimates of the probable response of the nonrespondents can be imputed (Groves, 2006; Singer, 2006).

Other approaches (e.g., Cochran, 1977) make no inferences about the probable responses of the nonrespondents, but rather determine what the outcome of the survey would be if all of the nonrespondents had answered one way or the other (e.g., the technique asks, for example, “What would be the result if *all* the nonrespondents said they planned to vote for the incumbent?”). This allows the researcher to draw boundaries on the percent of the total sample likely to vote for the incumbent. However, if the number of nonrespondents is sizable, this range can be so great that the survey has no practical utility. There are many examples of such correction

approaches in the sampling literature, because the problem is so common. Researchers interested in gaining the insights necessary to apply these solutions to a sampling problem, or to invent a solution that will work for a particular problem that might be encountered, are encouraged to consult this literature (e.g., Groves, 2006; Massey & Tourangeau, 2013; Singer, 2006). Adjusting for nonresponse is like the issue of dealing with missing data, to be discussed later in this chapter.

Applying Sampling Weights to Datasets

Due to the planned or unplanned oversampling or under-sampling of specific types of individuals (e.g., men, women, Democrats, Hispanic Americans, etc.) for a study, the researcher might obtain a final sample that is not representative of the population (Hahs-Vaughn & Lomax, 2006; Meier et al., 2013; Pfeffermann, 1996). Weights are applied to estimates to undo this nonrepresentative sampling. After the sample data are collected, weights are statistically derived to take the sampling design into account and to adjust for imbalances in proportions of various subgroups between the sample and the targeted population.

The most common type of weighting variable is called a *general sampling weight*, which is used to calibrate the demographic characteristics (e.g., age, gender, race, marital status, socioeconomic status) of the sample to a *known* population. Known population demographics are based on information obtained, for example, from the latest U.S. census or a previously conducted representative survey. This demographic adjustment will yield statistical estimates that are more representative of the general population. Additional types of weights also may be used to rectify other types of sampling problems. Participant nonresponse or attrition occurs when members of an identifiable subgroup are underrepresented in the sample because they refuse to participate at rates different from other groups. For instance, data may indicate that sampled participants over the age of 65 have a higher likelihood of refusal, perhaps because they are more distrustful of interviewers. Under-representation can be offset by judicious use of sampling weights.

Types of Survey Studies

Random Digit Dialing

Over the years, the use of the telephone as a data collection method for surveys has increased markedly. The technique of random digit dialing was developed for landline phone surveys, but over the years, landline usage has dropped precipitously. Using the telephone directory as a sampling frame is not common today. Indeed, Kennedy et al. (2018, p. 279) found that public opinion researchers have little to fear and potentially much to gain when landlines become a thing of the past.

Random digit dialing (RDD) has solved issues associated with the disappearance of the landline, but has some practical deficiencies of its own. The simplest random digit dial approach calls for the use of a random number generator to develop lists of telephone numbers. The ensuing numbers are the random sample of those to be called. However, most of the numbers generated in this manner are not in use, or fax numbers, or are not assigned to residential dwellings. Glasser and Metzger (1972) estimated that fewer than 20% of randomly generated phone numbers result in usable responses. When 80% of a survey researcher's calls are futile, the costs of the study are intolerable. Kunz and Fuchs (2012) have suggested ways of improving the odds of reaching eligible respondents using RDD, but the issue of nonresponse remains troublesome, especially given the practice of call screening.

The reporting of response rates also remains problematic in survey research. In some cases, response rates are nearly impossible to obtain from the published articles. More commonly, response rates can be determined in many ways, and this can have a material difference in judging the goodness of the sample. Martsolf et al. (2013) have discussed this issue and warned researchers and editors alike about the dangers inherent in non-standardized missing data reports for accurate interpretation of results.

Telephone versus Face-to-Face Interviews

Use of telephones to conduct surveys was seen as an improvement over the standard practice involving face-to-face interviews. The most obvious reason for the move to phone interviewing is that phone surveys can result in substantial savings of interviewer time and research funds. In fact, many large-scale national surveys today are done by telephone – the costs of doing otherwise would be prohibitive for most organizations. From the minute of its first use, researchers have debated the validity of telephone surveys relative to the formerly standard face-to-face model. In early attempts to validate the telephone survey approach, researchers compared their phone results with those obtained under the standard face-to-face conditions. If the results were similar, a vote was cast in favor of the telephone approach by virtue of its efficiency and cost savings.

Differences in results between the two survey modes often were taken as evidence of the failure of the phone survey to produce valid results, but this interpretation may be questionable. Consider a survey that deals with highly sensitive issues – respondents' use of illegal drugs, or other illegal acts, risky behaviors, and so on. In which context is the researcher likely to gain an honest answer? The quasi-anonymity of the phone conversation would seem to promote more honest answers to sensitive questions. Studies by Hoppe and colleagues (2000) and McAuliffe et al. (1998) revealed small differences in respondents' answers to questions in telephone versus face-to-face interviews, and when differences occurred, they suggested that the more valid answers were obtained over the phone, because higher rates of socially undesirable behaviors were reported. Boekeloo and colleagues (1998) found higher rates of reporting for both drug abuse and sexual encounters in their sample of early teenagers comparing telephone versus face-to-face interviews and concluded that the telephone survey was an acceptable means of obtaining valid data. This conclusion was reinforced by Meyer and colleagues, who successfully sampled gay and lesbian respondents using brief screening interviews in RDD samples (Meyer & Colten, 1999; Meyer et al., 2002).

Based on findings of this kind, Noble et al. (1998) touted the use of random digit dialing for large-scale policy research and argued the technique produced results that did not appear less valid than those derived from face-to-face interviews. Greenfield et al. (2000) agree and suggested the telephone survey approach appeared to produce results like the more costly interview formats. However, as Lavrakas (1993, 1998) suggested in his review of telephone survey methodology, caution should be exercised when using RDD. Because telephone service costs money, surveys making use of the phone, even RDD surveys, systematically under-sample the poor, the undereducated, the disenfranchised, and people of modest educational accomplishment. On some issues, this under-representation may produce biased results, and so care must be exercised when socio-demographic variations may be relevant to the issues under consideration.

Research that acknowledges the response rate difficulties now encountered in telephone interviewing has reconsidered use of address-based mail surveys (e.g., Link et al., 2008). Whether phone, live face-to-face, or mail surveys ultimately provide the highest quality data is an evolving question, whose answer changes with every passing decade. At one point, the mail survey reigned supreme, then the face-to-face approach captured scientists' favor, and then the random

digit dial model became popular. As discussed, these approaches have different strengths – economy, efficiency, likelihood of valid responses given the issue involved, etc. – but these strengths have evolved as time passed. The ultimate choice of technique is still under debate, but the choice of method will depend largely on available resources and the issues under study (Corso et al., 2022; Siebinga et al., 2022).

Internet Surveys

The development of computer-assisted self-administered interviews (CASAI) has made possible a new forum for survey research, namely, Web-based and e-mail surveys. Instead of mailing self-administered questionnaires to potential respondents, respondents can be contacted by e-mail or through the Web and directed to a site where they can complete a survey instrument online. This method of reaching a respondent sample has many advantages with respect to cost, time, and effort.

The major problem with electronic surveying is sampling representativeness. In most cases, the sampling frame for a Web-based survey is the population of people who have access to computers and feel comfortable using them regularly. In this sense, digital surveying approaches are in a similar position to that of telephone surveying in the 1950s, when many households still did not have telephone landlines. Even though the number of households with computer access is large and increasing, it is still true that people who have Internet access are likely to be more affluent, more educated, and younger than the population at large. Further, many Web-based surveys are conducted on a volunteer basis. A general announcement of the availability of an opportunity to participate in a social survey is posted on popular websites and interested respondents are instructed how to participate. Not surprisingly, the samples generated by such methods are highly self-selected and probably not typical of the population at large.

The use of sites at which potential participants may learn of research opportunities (e.g., Amazon's MTurk or CloudResearch, Prolific, or the Social Psychology Network) offer even more opportunities for survey researchers, and findings suggest that data collected on these sites are comparable to those obtained in face-to-face research (Buhrmester, Kwang, & Gosling, 2011; Paolacci, Chandler, & Ipeirotis, 2010). The broad utility of Internet- or Web-based surveys has encouraged its application across an enormous variety of studies, including study of trauma and mental health in military populations (Forkus et al., 2022), parents and children's interactions (e.g., Jensen Doss et al., 2022), revised post-traumatic stress measures (Petri et al., 2020), validity of crowdsourced longitudinal data on alcoholism (Strickland & Stoops, 2018), use of MTurk in studying eating disorders (Burnette et al., 2022).⁸ A general observation that seems to hold is that the validity of Internet-based surveys has not proven an issue in most research – and the rising number of such studies attests to its popularity and continued growth.

Missing Data

Each sampled person who chooses to participate in an investigation also must decide whether to answer each question in the survey instrument. This creates a different type of nonresponse problem from that discussed earlier. Respondents sometimes answer a survey only partially, skipping over certain questions they do not want to answer. Missing values on a question or questions bring into doubt the representativeness of the sample. Are the responses of those who answered a question representative of those who skipped it? Blank values are referred to as variable *missingness*. Although a variety of events may be responsible for missingness, especially problematic are sensitive questions concerning illegal activities, drug use, and sexual behavior.

Suppose a sizable number of sampled participants skipped an item that asked, “How many sexual partners have you had in your lifetime?” Any combination of these reasons is a possible explanation for respondents’ failure to answer the question: (a) They had too many partners and were too embarrassed to disclose their high number; (b) They had too few partners and were embarrassed to disclose their low number; (c) They couldn’t recall the exact number; (d) They didn’t understand the question (e.g., what constituted a “sexual partner?”); (e) No systematic reason, the skip was due to carelessness. Discovering the reasons underlying missing values is needed to deal with nonresponse bias. Classified according to the extent that nonresponse on one variable can be mathematically inferred from other measured variables, three types of missingness are possible: missing completely at random, missing at random, and not missing at random (Buhi et al., 2008; Roth, 1994; Schafer & Graham, 2002).

Missing completely at random (MCAR) occurs if the probability of missing responses in a variable is unrelated to any other variable in the study. Every person in the sample, that is, has an equal probability of responding or not responding to that question. Imagine taking an item with complete responses and then randomly deleting responses. The statistical estimate using all of the nonmissing responses would be sufficient to represent not only the missing responses but also all the responses had the entire sample answered the question. MCAR data might occur if, due to a printer malfunction, every other paper and pencil questionnaire accidentally omitted a particular question on half the questionnaires. The presumption is that no systematic difference exists in scores of the two versions to participants.

MCAR requires attrition analyses to ensure that the probability of the variable containing missing values is unrelated to (or uncorrelated with) any other variable. Suppose that 30% of all respondents did not answer the previous question about number of lifetime sexual partners. If no systematic difference in variable nonresponse was shown as a function of gender – an equal percentage of males and females skipped the question – this is initial evidence for MCAR. However, this alone is not sufficient. To fully test MCAR requires that we examine the variable with missing values against all other items in the survey to ensure independence.

Missing at random (MAR) occurs if the probability of missingness in a variable is *not* random, but its missingness may be fully explained by the other measured variables in the dataset. Unfortunately, the “random” in MAR is somewhat of a misnomer. It *does not* imply that the variable has an equal probability of whether respondents left it blank, but that the missingness is manageable as other measured variables in the dataset are entirely sufficient to account statistically for, and predict these empty values. Technically, the “random” refers to the deterministic notion that distributions of other variables could be used probabilistically to infer the missing values. Continuing with the scenario that revealed 30% of participants did not respond to a question on lifetime sexual partners, perhaps because of societal double standards in reporting, results might show that 50% of females but only 10% of males skipped this question. If the reason for the nonresponse is attributable entirely to measured variable(s) (e.g., respondents’ sex) found in the dataset, but not to any unmeasured variable(s) (e.g., social desirability, which was not assessed), then the requirement for MAR is satisfied.

Not missing at random (NMAR) occurs if a variable is not missing at random and its missingness is not explainable by the other measured variables in the dataset (but is explicable by other unmeasured variables). This is the least desirable result of a missingness analysis, because the missing values cannot be deduced statistically from other measured items in the dataset. NMAR implies that an unknown or unidentified event has caused the missing values. For instance, suppose the 30% of missing responses for the lifetime sexual partners question was attributable to participant social desirability bias. That is, 50% of those with high social desirability needs skipped the question, but only 10% of those low on social desirability needs skipped it.

As this survey focused exclusively on questions concerning sexual behaviors, respondents' need for social desirability was never measured, but this was the actual underlying reason for the differential rate of nonresponses. Had the culprit of social desirability bias been measured and identified as the factor that caused differences in response rates, the missingness variable would be reclassified as MAR, and its values could be imputed.

Possible solutions to missing data. To encourage higher response rates across questions, studies requiring a longer time commitment or interviews in people's homes usually offer some form of financial incentive. Another way to induce higher response rates is to assure confidentiality at the onset of the investigation, so that sampled individuals will not avoid questions because of fear that their privacy might be jeopardized. Keep in mind that NMAR, MCAR, and MAR apply to single variables, so it is possible that a dataset may contain different variables each with different forms of missingness. In theory, the assumption of MCAR or MAR is not always tenable, as important but unidentified reasons exist as to why people might skip or decline a question. Both MCAR and MAR are instances of variable nonresponses said to be largely *ignorable*, a statistical term for a manageable situation, because these blank values can be logically inferred from other measured variables by statistical imputation. If the missing data are assumed to be MCAR or MAR, the researcher may apply statistical software to automatically replace missing values with estimated values deduced from the information of other variables in the dataset. The most widely applied missing value imputation techniques are maximum likelihood and multiple imputation (see Mazza et al., 2022; Schafer & Graham, 2002; see also Baraldi & Enders, 2013; Grimm et al., 2014).

Conclusion

Surveys are ubiquitous in everyday life. Their design and the sampling rules that guide their application have been developed over the years in social science. The central feature of all good survey research has to do with the rules that guide respondent sampling. For many reasons, simple random samples are highly preferred, as their precision can be easily estimated with confidence. Simple random samples require a sampling frame and usually are improved by stratifying the sampled group based on theory-relevant variations. Cluster and multistage sampling, often found in large-scale studies of a population, are used when a sampling frame is not available. Non-random samples are used often, but their generalization to the underlying population is uncertain. Various tactics used to gather survey data – random digit dialing, face-to-face interviews, the Internet – are common in the survey world.

Notes

- 1 The U.S. census of the population is an incredibly costly endeavor, and in recent years, sophisticated social scientists have argued that a sample would be a cheaper and reliable (though unconstitutional) alternative to the census. Others have argued against the sampling alternative. This chapter illustrates the advantages and disadvantages of both positions.
- 2 With small samples, it is useful to modify the standard error estimate by a factor known as the *finite population correction*, or fpc, as follows:

$$\text{S.E.} = \sqrt{\frac{(1-f)s^2}{n}}$$

where f = the sampling fraction, i.e., the proportion of the total population included in the sample.

- 3 An issue in the definition of simple random sampling is whether sampling is conducted with or without replacement. Following Kalton (2020, p. 10), we take this term to refer to sampling without replacement; that is, once a unit is included in the sample, he, she, or it is not returned to the population

- and, thus, cannot be chosen again. Sometimes, the term simple random sampling without replacement is used to describe this form of sampling, and the term simple random sampling with replacement is used to describe the equivalent technique in which the units sampled at one draw are returned to the population before the next draw is made. In practice, almost all sampling is done without replacement.
- 4 Random number tables can be found in the appendix sections of most statistics textbooks, and today can be generated on most personal computers with almost any basic statistical software or Internet access.
 - 5 The overweighing of the smaller group is reversed in the later statistical analysis of the data, so its actual strength is represented accurately. Overweighing is used to ensure a sufficient number of respondents to ensure a more representative sampling of the small group.
 - 6 Although we have limited our discussion to a simple two-phase sampling process, the number of ancillary surveys that can be undertaken is limited only by cost considerations -- and the patience of the respondents in the sample. When more than two sampling phases are used, the approach is termed multi-phase sampling.
 - 7 That is, the estimate is within 10% of the population percentage, with a 95% probability.
 - 8 Note that this study met with considerable constructive suggestions (e.g., Moeck et al., 2022; Vogel et al., 2022).

References

- Baraldi, A. N., & Enders, C. K. (2013). Missing data methods. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods: Statistical analysis* (Vol. 2, pp. 635–664). Oxford University Press.
- Blalock, H. M. (1985). *Causal models in panel and experimental designs*. Aldine.
- Blumberg, S. J., & Luke, J. V. (December, 2021). Wireless substitution: Early release of estimates from the National Health Interview Survey, January–June 2021. National Center for Health Statistics. Retrieved from <https://www.cdc.gov/nchs/data/nhis/earlyrelease/wireless202111.pdf>.
- Boekeloo, B. O., Schamus, L. A., Simmens, S. J., & Cheng, T. L. (1998). Ability to measure sensitive adolescent behaviors via telephone. *American Journal of Preventive Medicine*, 14, 209–216.
- Buhi, E. R., Goodson, P., & Neilands, T. B. (2008). Out of sight, not out of mind: Strategies of handling missing data. *American Journal of Health Behavior*, 32, 83–92.
- Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source for inexpensive, yet high-quality data? Amazon's Mechanical Turk: A new source of inexpensive, yet high quality, data? *Perspectives on Psychological Science*, 6, 3–5.
- Burnette, C. B., Luzier, J. L., Bennett, B. L., Weisenmuller, C. M., Kerr, P., Martin, S., Keener, J., & Calderwood, L. (2022). Concerns and recommendations for using Amazon MTurk for eating disorder research. *International Journal of Eating Disorders*, 55(2), 263–272.
- Callegaro, M., McCutcheon, A. L., & Ludwig, J. (2010). Who's calling? The impact of caller ID on telephone survey response. *Field Methods*, 22, 175–191.
- Chung, S., & Seo, J. (2007). A study on posttraumatic stress disorder among North Korean defectors and their social adjustment in South Korean. *Journal of Loss and Trauma*, 12, 365–382.
- Church, A. H. (1993). Estimating the effect of incentives on mail survey response rates: A meta-analysis. *Public Opinion Quarterly*, 57, 62–79.
- Cochran, W. G. (1977). *Sampling techniques*. Wiley.
- Cook, D. A., Wittich, C. M., Daniels, W. L., West, C. P., Harris, A. M., & Beebe, T. J. (2016). Incentive and reminder strategies to improve response rate for Internet-based physician surveys: A randomized experiment. *Journal of Medical Internet Research*, 18(9), 107–117.
- Cooper, S. L. (1964). Random sampling by telephone -- An improved method. *Journal of Marketing Research*, 1, 45–48.
- Corso, M., Cancelliere, C., Mior, S., Salmi, L. R., Cedraschi, C., Nordin, M., Sci, D. M., Taylor-Vaisey, A., & Côté, P. (2022). Are nonpharmacologic interventions delivered through synchronous telehealth as effective and safe as in-person interventions for the management of patients with nonacute musculoskeletal conditions? A systematic rapid review. *Archives of Physical Medicine and Rehabilitation*, 103(1), 145–154.
- Crano, W. D., Siegel, J. T., Alvaro, E. M., Lac, A., & Hemovich, V. (2008). The at-risk adolescent marijuana nonuser: Expanding the standard distinction. *Prevention Science*, 9, 129–137.

- Cycyota, C. S., & Harrison, D. A. (2002). Enhancing survey response rates at the executive level: Are employee- or consumer-level techniques effective? *Journal of Management*, 28(2), 151–176.
- Forkus, S. R., Contractor, A. A., Goncharenko, S., Goldstein, S. C., & Weiss, N. H. (2022, May 19). Online Crowdsourcing to Study Trauma and Mental Health Symptoms in Military Populations: A Case for Amazon's Mechanical Turk (MTurk) Platform. *Psychological Trauma: Theory, Research, Practice, and Policy*. Advance online publication. <http://dx.doi.org/10.1037/tra0001235.supp> (Supplemental).
- Gaeddert, L. A., Schneider, A. L., Miller, C. N., Monteith, L. L., Brenner, L. A., Katon, J., & Hoffmire, C. A. (2020). Recruitment of women veterans into suicide prevention research: Improving response rates with enhanced recruitment materials and multiple survey modalities. *Research in Nursing & Health*, 43(5), 538–547.
- Glasser, G. J., & Metzger, G. D. (1972). Random-digit dialing as a method of telephone sampling. *Journal of Marketing Research*, 9, 59–64.
- Greenfield, T. K., Midanik, L. T., & Rogers, J. D. (2000). Effects of telephone versus face-to-face interview modes on reports of alcohol consumption. *Addiction*, 95, 277–284.
- Griffin, J. M., Simon, A. B., Hulbert, H., Stevenson, J., Grill, J. P., Noorbaloochi, S., & Partin, M. R. (2011). A comparison of small monetary incentives to convert survey non-respondents: Randomized control trial. *BMC Medical Research Methodology*, 11, 81.
- Grimm, K. J., Ram, N., Shiyko, M. P., & Lo, L. L. (2014). Simulation study of the ability of growth mixture models to uncover growth heterogeneity. In J. J. McArdle, & G. Ritschard (Eds.), *Contemporary issues in exploratory data mining in the behavioral sciences* (pp. 172–189). New York, NY: Routledge.
- Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 70, 646–675.
- Hadland, S. E., Marshall, B., Kerr, T., Qi, J., Montaner, J. S., & Wood, E. (2012). Suicide and history of childhood trauma among street youth. *Journal of Affective Disorders*, 136, 377–380.
- Hahs-Vaughn, D. L., & Lomax, R. G. (2006). Utilization of sample weights in a single-level structural equation modelling. *The Journal of Experimental Education*, 74, 163–190.
- Hazan, C., & Shaver, P. S. (1987). Romantic love conceptualized as an attachment process. *Journal of Personality and Social Psychology*, 52, 511–524.
- Henry, G. T. (1990). *Practical sampling*. Sage.
- Homish, G. G., & Leonard, K. E. (2009). Testing methodologies to recruit adult drug-using couples. *Addictive Behaviors*, 34, 96–99.
- Hoppe, M. J., Gillmore, M. R., Valadez, D. L., Civic, D., Hartway, J., & Morrison, D. M. (2000). The relative costs and benefits of telephone interviews versus self-administered diaries for daily data collection. *Evaluation Review*, 24, 102–116.
- Jacoby, L. H., Young, B., & Watt, J. (2008). Public disclosure in research with exception from informed consent: The use of survey methods to assess its effectiveness. *Journal of Empirical Research on Human Research Ethics*, 3, 79–87.
- Jensen-Doss, A., Patel, Z. S., Casline, E., Mora Ringle, V. A., & Timpano, K. R. (2022). Using Mechanical Turk to study parents and children: An examination of data quality and representativeness. *Journal of Clinical Child and Adolescent Psychology*, 51(4), 428–442.
- Kalton, G. (2020). *Introduction to survey sampling* (2nd Ed.). Newbury Park, CA: Sage.
- Kempf, A. M., & Remington, P. L. (2007). New challenges for telephone survey research in the twenty-first century. *Annual Review of Public Health*, 28, 113–126.
- Kennedy, C., McGeeney, K., Keeter, S., Patten, E., Perrin, A., Lee, A., & Best, J. (2018). Implications of moving public opinion surveys to a single-frame cell-phone random-digit-dial design. *Public Opinion Quarterly*, 82(2), 279–299.
- Kish, L. A. (1965). *Survey sampling*. Wiley.
- Kunz, T., & Fuchs, M. (2012). Improving RDD cell phone samples: Evaluation of different pre-call validation methods. *Journal of Official Statistics*, 28, 373–394.
- Lavrakas, P. J. (1993). Telephone survey methods: Sampling, selection, and supervision. In *Applied social research methods series* (2nd ed., Vol. 7). Newbury Park, CA: Sage.
- Lavrakas, P. J. (1998). Methods for sampling and interviewing in telephone surveys. In L. Bickman, & D. J. Rog (Eds.), *Handbook of applied social research methods* (pp. 429–472). Sage.
- Levy, P. S., & Lemeshow, S. (2008). *Sampling of populations: Methods and applications*. Wiley.

- Link, M. W., Battaglia, M. P., Frankel, M. R., Osborn, L., & Mokdad, A. H. (2008). A comparison of address-based sampling (ABS) versus random-digit dialing (RDD) for general population surveys. *Public Opinion Quarterly*, 72, 6–27.
- Link, M. W., & Mokdad, A. (2005). Leaving answering machine messages: Do they increase response rates for RDD Surveys? *International Journal of Public Opinion Research*, 17, 226–250.
- Massey, D. S., & Tourangeau, R. (2013). New challenges to social measurement. *The Annals of the American Academy of Political and Social Science*, 645, 6–22.
- Martsoff, G. R., Schofield, R. E., Johnson, D. R., & Scanlon, D. P. (2013). Editors and researchers beware: Calculating response rates in random digit dial health surveys. *Health Services Research*, 48(2, Pt1), 665–676.
- Mazza, G. L., & Enders, C. K. (2014). Missing data analysis. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd ed., pp. 627–652), Cambridge University Press.
- Mazza, G. L., Petersen, M. M., Ginos, B., Langlais, B. T., Heon, N., Gounder, M. M., Mahoney, M. R., Zoroufy, A. J., Schwartz, G. K., Rogak, L. J., Thanarajasingam, G., Basch, E., & Dueck, A. C. (2022). Missing data strategies for the Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE) in Alliance A091105 and COMET-2. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation*, 31(4), 1069–1080.
- McAuliffe, W. E., Geller, S., LaBrie, R., Paletz, S., & Fournier, E. (1998). Are telephone surveys suitable for studying substance abuse? Cost, administration, coverage and response rate issues. *Journal of Drug Issues*, 28, 455–481.
- Meier, P. S., Meng, Y., Holmes, J., Baumberg, B., Purshouse, R., Hill-McManus, D., & Brennan, A. (2013). Adjusting for unrecorded consumption in survey and per capita sales data: Quantification of impact on gender- and age-specific alcohol-attributable fractions for oral and pharyngeal cancers in Great Britain. *Alcohol and Alcoholism*, 48, 241–249.
- Meyer, I. H., & Colten, M. E. (1999). Sampling gay men: Random digit dialing versus sources in the gay community. *Journal of Homosexuality*, 37, 99–110.
- Meyer, I. H., Rossano, L., Ellis, J. M., & Bradford, J. (2002). A brief telephone interview to identify lesbian and bisexual women in random digit dialing sampling. *Journal of Sex Research*, 39, 139–144.
- Mickelson, K. D., Kessler, R. C., & Shaver, P. R. (1997). Adult attachment in a nationally representative sample. *Journal of Personality and Social Psychology*, 73, 1092–1106.
- Moeck, E. K., Bridgland, V. M. E., & Takarangi, M. K. T. (2022). Food for thought: Commentary on Burnette et al (2021). ‘Concerns and recommendations for using Amazon MTurk for eating disorder research’. *International Journal of Eating Disorders*, 55(2), 282–284.
- Moser, C. A., & Kalton, G. (1972). *Survey methods in social investigation* (2nd ed.). Basic Books.
- Noble, I., Moon, N., & McVey, D. (1998). ‘Bringing it all back home’-- Using RDD telephone methods for large-scale social policy and opinion research in the UK. *Journal of the Market Research Society*, 40, 93–120.
- Oldendick, R. W., & Link, M. W. (1994). The answering machine generation: Who are they and what problem do they pose for survey research? *Public Opinion Quarterly*, 58, 264–273.
- Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5, 411–419.
- Patrick, M. E., Couper, M. P., Parks, M. J., Laetz, V., & Schulenberg, J. E. (2021). Comparison of a web-push survey research protocol with a mailed paper and pencil protocol in the Monitoring the Future panel survey. *Addiction*, 116(1), 191–199.
- Petri, J. M., Weathers, F. W., Witte, T. K., & Silverstein, M. W. (2020). The detailed assessment of posttraumatic stress–second edition (daps-2): Initial psychometric evaluation in an MTurk-recruited, trauma-exposed community sample. *Assessment*, 27(6), 1116–1127.
- Pfeffermann, D. (1996). The use of sampling weights for survey data analysis. *Statistical Methods in Medical Research*, 5, 239–261.
- Rahn, W. M., Krosnick, J. A., & Breuning, M. (1994). Rationalization and derivation processes in survey studies of political candidate evaluation. *American Journal of Political Science*, 38, 582–600.
- Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology*, 47, 537–560.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147–177.

- Siebinga, V. Y., Driever, E. M., Stiggelbout, A. M., & Brand, P. L. P. (2022). Shared decision making, patient-centered communication and patient satisfaction – A cross-sectional analysis. *Patient Education and Counseling*, 105(7), 2145–2150.
- Singer, E. (2006). Introduction: Nonresponse bias in household surveys. *Public Opinion Quarterly*, 70, 637–645.
- Strickland, J. C., & Stoops, W. W. (2018). Feasibility, acceptability, and validity of crowdsourcing for collecting longitudinal alcohol use data. *Journal of the Experimental Analysis of Behavior*, 110(1), 136–153.
- Troldahl, V. C., & Carter, R. E., Jr. (1964). Random selection of respondents within households in phone surveys. *Journal of Marketing Research*, 1, 71–76.
- Trussel, N., & Lavrakas, P. J. (2004). The influence of incremental increases in token cash incentives on mail survey response. *Public Opinion Quarterly*, 68, 349–367.
- Tuckel, P., & O'Neill, H. (2002). The vanishing respondent in telephone surveys. *Journal of Advertising Research*, 42, 26–48.
- Vogel, M., Krüger, J., & Junne, F. (2022). Eating disorder related research using Amazon Mechanical Turk (MTurk): Friend or foe?: Commentary on Burnette et al (2021). *International Journal of Eating Disorders*, 55(2), 285–287.
- Xu, M., Bates, B. J., & Schweitzer, J. C. (1993). The impact of messages on survey participation in answering machine households. *Public Opinion Quarterly*, 57(2), 232–237.

13

SYSTEMATIC OBSERVATIONAL METHODS

In the previous section of this book, we were concerned with general strategies used in the design and implementation of different types of research studies. We now turn attention to methods of assessment and measurement used to collect the data that constitute the output of our research operations. We begin this section with a general consideration of systematic observational methods. This is appropriate, because all the assessment or measurement techniques to be presented require some form of systematic observation. Although all science is bound fundamentally to observation, the term “systematic observational methods” in our view has come to refer to a diverse set of techniques that are employed to study behavior that:

- (Usually) occurs outside the formal boundaries of the laboratory
- (Usually) is naturally instigated, i.e., does not make use of a controlled experimental treatment
- (Usually) places few restrictions on the allowable responses of the persons under observation
- (Usually) focuses on observable behaviors rather than internal, cognitive processes
- (Usually) entails a replicable system of codifying observed events

This last requirement for observational data separates *systematic* observational techniques from unsystematic research approaches that may play a role in hypothesis generation or theory development, but do not constitute formal data collection.

The first four bulleted points of our description of observational methods suggest both the range of research tactics that can be legitimately described as observational, and the variations that characterize the uses of this general approach to accumulate scientifically informative data. Systematic observation may be a feature of the experimental laboratory, where behaviors are constrained by the researcher, but in this chapter we focus our discussion principally on the use of observation in *natural* settings, where participants’ repertoire of behaviors are usually not constrained.

Three Aspects of Naturalism

The naturalistic character of data obtained via systematic observational methods is the most valued feature of this type of research – it is the major reason why research is conducted outside the

confines of the laboratory. In his discussion of the dimensions of field research, Tunnell (1977) observed that there was some confusion about how the naturalness of a research enterprise should be conceptualized. The criterion of *naturalness*, he observed, could be applied to the *behaviors* being studied, the *treatments* that are (or are not) applied, and the *setting* in which the research is conducted. Furthermore, these facets could be combined in any number of ways, and the combinations would reflect the extent to which the total study could be judged as more or less naturalistic.

At the extreme end of “unnaturalness,” we have studies that constrain actions or require behavior that is not a usual part of a person’s behavioral repertoire, make use of strong manipulations, and occur in unusual contexts. At the other end of the naturalness dimension is research that places no constraints on participants’ behaviors, that does not impinge on the environment, and that occurs in natural settings. In this view, all scientific research involves observation of a dependent variable, but the behaviors, impingements, and settings on which the research is focused can vary widely, and Tunnell’s (1977) system helps us situate research along the three dimensions of naturalness.

Natural Behavior

A primary goal of almost all observational research is to study naturally instigated behavior. As discussed (e.g., Chapter 7), impressive gains in generalizability can be realized if the behavior under study is naturally instigated and not made in response to the demands of the research situation or experimenter. Behavior is considered natural to the extent that it is an existing part of the individual’s response repertoire, and not enacted to meet the demands of the study, and is not self-conscious (i.e., it is produced without the actor’s self-conscious awareness of being the object of scientific scrutiny).

Many field studies are undertaken to transfer the control of the laboratory to naturalistic field settings, allowing study of behavior that is generated by the participant, not the researcher. Such studies avoid restricting the range of possible reactions specified in advance by the experimenter. In a clever field experiment, Moriarty (1975) induced individuals sunbathing at a crowded beach to feel responsible for the welfare of another by having a confederate ask a fellow sunbather (the participant) to watch his radio while he went “to the boardwalk for a few minutes.” Soon thereafter, another confederate approached the empty blanket, picked up the radio, and, if not stopped by the participant, ran off with it.

Participants’ responses to the apparent theft of the property constituted the dependent measure. In this instance, participants’ behaviors were natural, that is, unconstrained and unself-conscious (in the sense that most probably did not think they were part of a psychological study). Their possible responses ranged from physically restraining the thief, following him, calling for help, leaving the scene, or ignoring the episode entirely. The naturalistic behavioral responses obtained in this research were used to study the relation between responsibility on prosocial behavior, or bystander intervention, and the findings of this research were judged by many to be more applicable to natural human behavior than those of earlier investigations that had been conducted in more artificial laboratory settings.¹ In the naturalistic approach, emphasis is placed on examining people’s observable behavior rather than beliefs and thoughts.

Natural Event

In Tunnell’s (1977, p. 428) view, a natural treatment is a “naturally occurring, discrete event... that the subject would have experienced...with or without the presence of a researcher.” By

this definition, Moriarty's (1975) treatment (a staged theft of a radio) was not natural, because it required the researcher's active involvement. Although we are in general agreement with Tunnell's classification system, we find his natural treatment definition overly restrictive, and inconsistent with the general intention of fostering more and better field research. In our view, a treatment (an externally controlled or created event) can be considered natural even if produced by the actions of an experimenter if (1) it plausibly could have occurred without an experimenter's intervention, and (2) the participant is unaware of the fact that it did not. Thus, we would consider Moriarty's context and treatment "natural," even though it was introduced by the researcher and, thus, decidedly "unnatural." In our view, the experimental manipulation of the situation did not destroy its naturalness – the theft of people's belongings on public beaches does occur – and the participants' reactions suggest strongly that they did not suspect that they were involved in a scientific investigation.²

Natural Setting

Generally, in the social sciences a "naturalistic observational investigation" refers to a study that has been conducted outside the laboratory. Whether the researcher has intervened in the setting is irrelevant, so long as the research participants do not perceive it to have been established for research purposes. As in our discussion of natural treatments, in our view respondents' perceptions of the setting, not the actions of the researcher, define the naturalness of the setting. In this sense, any setting can be considered natural if the respondent does not think it has been constructed for the purpose of research. Thus, a college classroom can be a natural setting (at least for college students). Moriarty's beach scene is an obvious (good) example of a natural setting. It is reasonable to assume that people were unaware that the setting had been modified to allow for a systematic study of the effects of the specification of responsibility on helping behavior. The setting contributed to the credibility of the treatment. This last point bears emphasis: even though we might take great care in the development of hypotheses, manipulations, measures, etc., we sometimes are lax in our choice of the setting in which the observations occur. This is a mistake, because an ill-chosen setting can defeat even the most well-designed study by rendering data collection difficult or impossible (Weick, 1985).

It is important to understand that all three of Tunnell's dimensions of naturalness interact, and thus, could produce a situation in which the observed responses are very different from those we expect. In a study similar to Moriarty's (1975), Howard and Crano (1974) staged a theft by a confederate of books owned by another person (also a confederate) to examine people's reactions. In our terms, the treatment, setting, and behaviors were completely naturalistic. The (natural) settings for this research were the university library, the student union, and a popular on-campus grille. The treatments (the behaviors of the victim and the thief) were well rehearsed and generated no suspicion. And, of course, the behaviors of the respondents in reacting to the theft were unconstrained. Even so, there was surprising (and unexpected) behavioral variation in the degree to which bystanders were willing to intervene (by stopping or identifying the book thief) in the three settings. Relatively little help was provided the victim in the library, more in the lounge, and most in the grille. The authors reasoned that different norms dictated appropriate behavior in the settings and were responsible for the variations in helping that were observed. Such norms can influence behavior powerfully even when events call for a radical departure from the usual or prescribed forms of action. Though unexpected, these findings were more credible because of the naturalistic circumstances in which they were observed.

Whereas Tunnell's (1977) three-dimensional classification used to judge the extent that research is naturalistic is useful, it is important to realize that many other features involving

the design of the study play a role in determining the extent to which naturalness is achieved. Two of the most important of these primary design considerations are the extent to which the observer is involved in the activities under investigation, and the type and form of coding system chosen to summarize the behaviors that are the focus of the study. These two factors lead to a series of more specific tactical considerations that guide the conduct of any observational study.

Observer Involvement in the Naturalistic Setting

Participatory Observation

The degree to which an observer interacts with the individuals under study is one of the most important determinants of the procedures of the study, the quality of the data that will be collected, and the interpretation of obtained results. Observer interaction with the observed can vary tremendously, from complete engagement with those observed in the situation to removal in both time and space from those being studied. The technique calling for the greatest intimacy between observer and observed, **participatory observation** (or participant observation), is an “intense social interaction between researchers and participants in the milieu of the latter, during which time data, in the form of field notes, are unobtrusively and systematically collected” (Bogden, 1972, p. 3; also see Østerlund et al., 2016; Watt & Jones, 2010).

Participatory observation is most widely used in sociology, ethnography, and cultural anthropology, and is viewed by many in these fields as an indispensable feature of those disciplines. Methods involving participatory observation have given rise to considerable controversy over the years, probably because the method allows the researcher so much freedom in defining his or her appropriate realm of action and demands so little systematization of the observation techniques that are employed. Indeed, Williamson and Karp (1977) observed that one of the most remarkable features of this approach was the *lack* of agreed-upon rules or standard operating procedures used to guide its use. Confirming this observation, the venues and topics in which participatory observation approaches have been employed are vast, ranging from studies of restraint practices in an acute care hospital (Thomann et al., 2022), to “FBI defined ‘hybrid’ gangs” (Przemieniecki et al., 2020), go-go dancers in the Philippines (Ratliff, 1999), and communication in groups on the Internet (Polifroni et al., 2001).

DeWalt and DeWalt (2010) provided a good working definition of participatory observation as

a method in which a researcher takes part in the daily activities, rituals, interactions, and events of a group of people as one of the means of learning the implicit and tacit aspects of their life routines and their culture.

The participatory observer, in other words, becomes a part of the everyday life of the observed, while also gathering data. The observer is both a part of the ongoing behavior of the group, and so, almost inevitably affects it. This feature of participatory observation is responsible for much of the controversy that surrounds the approach. It is an inevitability of the method, its greatest strength and its greatest potential weakness. We return to this issue when discussing the relative advantages and disadvantages of participatory and non-participatory techniques.

The term participatory observation, then, broadly describes the general research process in which an observer observes and is embedded within the observational context. The observer passes as or becomes a member of a group and uses this insider status to gather data. Often, participatory observation entails the simultaneous collection and analysis of data (Atkinson &

Hammersley, 1994; Watt & Jones, 2010), because in this technique, the processes of hypothesis generation and hypothesis testing often occur almost simultaneously.

Gaining entry. To make best use of this technique, the observer first must gain entrance into the group under study. A good example of some of the possible dangers and distortions involved in this entry process was provided in a now classic study of the *Seekers* by Festinger et al. (1956). The *Seekers* was a cult that claimed to have had contact with extra-terrestrial beings, the *Guardians*. These benevolent space people had presumably informed the cult leader (Dorothy Martin) that a major flood was soon to inundate the northern hemisphere. The *Seekers* believed the *Guardians* would eventually send a flying saucer at the last minute to transport them out of harm's way. The cult was not large, but belief in the prophecy was strong among its members.

Not sharing the faith, Festinger and colleagues (1956) decided this group provided a good opportunity to study the effects of disconfirmation of a strong expectation on reinforcing people's future behavior. The best way to accomplish this investigation, they reasoned, would be to observe the cult from the inside, and thus obtain an intimate view of their proceedings. The *Seekers* were not a proselytizing group – membership was by invitation only. As the authors admitted,

our basic problems were...obtaining entree for a sufficient number of observers to provide the needed coverage of members' activities and keeping at an absolute minimum any influence which these observers might have on the beliefs and actions of members of the group. We tried to be nondirective, sympathetic listeners, passive participants who were inquisitive and eager to learn whatever others might want to tell us.

(p. 237)

To gain entry into the cult, members of Festinger's team hid their academic credentials, approached the *Seekers*, and claimed "psychic experiences" of the type calculated to interest the group's members. The psychic experiences were constructed so that they could be interpreted as consistent with the *Seekers*' system of beliefs. The researchers recognized the danger that these stories might have interfered with the natural equilibrium of the group but was judged necessary because of the *Seekers*' sensitivity to ridicule:

Unhappily, [the ruse] had been too successful, for, in our effort to tailor a story to fit the beliefs of the members of the group, and thus gain their approval for our observers, we had done too well. We had unintentionally reinforced their beliefs that the *Guardians* were watching over humanity and were 'sending' chosen people for special instruction about the cataclysm and the belief system.

(p. 241)

Unfortunately, these fears were realized. The *Seekers* interpreted the stories concocted by the new members as supporting their beliefs.

The necessary entry-gaining actions of Festinger's observers illustrate a point that should be recognized in all studies making use of participatory observer techniques; that is, in almost all closed groups, the problem of entry assumes great importance. The actions the non-member performs in gaining admittance to the group can affect the ongoing actions of the group and the quality of the data. Thus, in attempting to join a group for the purposes of participatory observation, the most unobtrusive means possible should be used, so the natural group situation remains so. Failing that, a detailed account of possible interference resulting from the entry

process should be included in descriptions of the research, as was in Festinger et al. (1956). This description will not solve the interference problem, but at least will alert the reader to possible distortions that might have occurred.

A different approach to entry (and a different set of problems) is evident in Thorne's (1988) study of the draft resistance movement during the Vietnam War. Thorne first joined the anti-draft movement, and later decided to use her experiences as a source of data for her thesis. To avoid deceiving her peers, Thorne discussed her plans, which were met with responses that ranged from "hostility to mild tolerance" (p. 134). Whereas Thorne avoided deceiving those who shared her political convictions, her openness may have altered the fundamental nature of the interactions that ensued *after* she disclosed her research plans. By disclosing her intentions, she was able to maintain truthfulness to a much greater degree than the participatory observers of Festinger et al. (1956). It could be argued, however, that her disclosure affected the naturalness of the context just as strongly as Festinger's observers had.

These two extreme examples do not offer a clear solution to the problem of entry. Complete openness as to motive and approach is certainly more ethically defensible, but in no way does it solve the problem of the observer affecting the observed. Complete nondisclosure of techniques and motivation, as exemplified by Festinger et al. (1956), would appear to solve the problem of group members reacting unnaturally to one of their group; unfortunately, the way observers gained entry, and the effects of their actions while in the group, can take a toll on the quality and credibility of the behavioral observations that are made.

Going native. The entry problem is only one of the difficulties encountered when using participatory observation. Just as there is a large literature in anthropology, psychology, ethnology, and sociology devoted to the solution of entry problems, so too have social scientists been concerned with the problem of "going native." Overidentification with the observed group tends to blind the observer to many relevant aspects of the total situation and draws attention to those events perceived to be of interest to the group, which may or may not be scientifically worthwhile. Given Thorne's (1988) identification with the draft resistance movement, one might question the quality of the data her research produced. Could she be an unbiased and nonjudgmental observer, given her political commitments?

A hypothetical example will help to illustrate this point. Suppose a pair of agnostic researchers were interested in the patterns of interaction that occurred during prayer services of a Pentecostal church; over time, observing the group, the researchers became so deeply impressed by the fervor of the members that they participated in the church's activities, ultimately becoming church members. Would their accounts of the group's activities and practices prove of interest? Possibly. A more difficult question concerns the scientific value of the work. The fact of the researchers' conversion does not necessarily diminish their scientific expertise. It is the way this expertise is employed that gives cause for concern.

We do not suggest that researchers cannot become close to those they observe (see Mehan & Wood, 1975; Thorne, 1988), but rather that the ultimate purpose of the engagement, the systematic collection of reliable information, should be kept in mind. It is the overly complete identification with the group under study that can lead to failure to integrate new with known information (a kind of scientific tunnel vision), and discounting information that casts the group in a poor light, that proves so troublesome when the observer "goes native."

Non-Participatory and Partial Participatory Observation

By restricting the degree of participant observation in their research, investigators can simultaneously control the problems both of entry and of going native. This solution has generated

studies in which observers were, for the most part, unseen by the observed, their existence oftentimes not even suspected.

Less remote from the actual events under consideration are investigations of ongoing group or individual activity in which observers are partially or totally concealed. Total concealment requires a highly structured observational setting. It is essential the group being studied remain in the setting. Attempts at completely concealed, nonparticipant observation have resulted in a variety of research operations that are sometimes interesting, often amusing, and almost always ethically questionable. Henle and Hubbell's (1938) observers, for example, hid under beds in college dormitories to collect their data. We do not recommend this strategy. Less dangerous, but still ethically questionable, was research by Carlson et al. (1936) in which observers eavesdropped on people's conversations in theater lobbies. The list could be extended indefinitely, but concealed observation is usually an ethically questionable operation. Such research techniques should be considered only after all other options have proved infeasible.

A variant on non-participatory observation makes use of structured diaries and time sampling (DeLongis et al., 1992; Leigh, 1993; Robinson & Goodbey, 1997), in which respondents complete a set of questionnaires regarding their mood, their perceptions of interactions with partners, thoughts, actions, etc., at time intervals specified by the researcher (say, daily, or at randomly determined intervals). In effect, this method relies on self-observation by the participants under study. Robinson and Goodbey (1997), for example, asked participants to monitor their time allocations (leisure vs. obligatory) over the course of the day, and Leigh (1993) had participants note their alcohol consumption and sexual activity at the end of each day. The "observations" made under these circumstances are of the self-report, retrospective, variety, but the naturalism of the setting in which the data collection takes place, the natural behavior under study, the short duration between action and recording, and the natural events that occur place this approach among the naturalistic methods employing non-participatory observation.

A common example of *partial* participatory observation is found in studies of the patient-therapist relationship. In research of this type, an observer (often, a student in training) attends the therapeutic hour, but strictly as an observer – the observer does not interact in any way with either patient or therapist. Another even more common variant on this theme involves filming the therapeutic interaction for later study. Often, both therapist and patient know of the ongoing observation, but it is felt that this form of partial participation is less intrusive than that in which the observer is physically present. However, as might be expected, research indicates that people who know they are being watched tend to emit more socially desirable behaviors in either context. Roberts and Renzaglia (1965), for example, showed that the mere presence of a microphone had a measurable effect on the behaviors of both patient and therapist. Zegiob et al. (1975) found people were more likely to play positively with children when being observed. These studies suggest that the mere fact of observation can bias the observational record. Despite this, today the use of audio or video recordings in social research is becoming increasingly common.

These approaches have the advantage of allowing the researcher to examine and re-examine the behaviors of interest, at a time of his or her choosing. The capability of "taking a second look" to resolve possible misinterpretation of the observations usually enhances the reliability of any behavioral coding. The point at which techniques of this sort fall on the participant-nonparticipant continuum depends on the obtrusiveness of the recording device. If participants are unaware of the observation, the research can be judged by the investigator as non-participatory (if, perhaps, unethical); if respondents are aware of the taping, and the recording machinery is an obvious part of the setting, then as Roberts and Renzaglia (1965) demonstrated, the degree of observer interference can be as severe as it would have been had the observer been physically present.

To avoid reliance on retrospective reports of behavior, researchers may use an **experience sampling method** (ESM), otherwise known as ecological momentary assessments, in which respondents carry a computer tablet, smart phone, or other electronic reminder device; when the device sounds, each respondent completes questions about what they are doing, thinking, feeling at that moment, etc. (Bolger & Laurenceau, 2021; Silvia & Cotter, 2021). Alternately (see Iida et al., 2012), participants may be asked to complete a measure at a specified time (or times) each day (Gable et al., 2000). Usually, data are collected continually for days, weeks, or even months. If cell phone is used, it usually is programmed to contact respondents once in every given block of time (say, once every two hours or every day), although a random time schedule also may be used (Kubey et al., 1996; Larson & Csikszentmihalyi, 1983). ESM is a time-sampling technique intended to obtain slices of life events across people's everyday behaviors and activities.

The value of the ESM diary approach is that it “catches” respondents in their natural milieu, and thus allows the researcher to associate the outcomes of various natural events with respondents' subsequent moods, intentions, behaviors, etc. If data are drawn on a random schedule, they would appear less susceptible to the cyclical biases that might occur if the diaries were always completed at the same times of day for a month. If participants were simply asked to complete measures at the end of each night, they may exhibit a tendency to place a greater emphasis on temporally near events, such as experiences occurring that evening, or might be more susceptible to distortions in accurately recalling behaviors and mood states earlier that day. This daily approach seems less challenging than one-time retrospective surveys in which participants are asked to notate their attitudes and behaviors of the more distant past. Data collected from the ESM are usually estimated with multilevel models (see Chapter 11).³

The ESM and other diary methods can be used creatively to study a variety of issues. Larson, Csikszentmihalyi, and their colleagues, for example, used this method to document the negative effects of solitude on people's feelings, and the generally positive after-effects of no longer being alone (Larson & Csikszentmihalyi, 1978, 1980; Larson et al., 1980). It is difficult to envision how solitude could be studied credibly in a laboratory context. Nezlek et al. (1983) employed a structured diary approach to study such issues as the initiation and personal control of social encounters. During a two-week period, participants were instructed to record responses to questions asking about every social encounter lasting at least ten minutes. These studies provide valuable real-time insights into effects that everyday events have on people's natural behaviors. As such, these kinds of diary approaches have become increasingly popular (Itzhakov & Reis, 2021; Reis & Gable, 2000).

The ESM has been used in a variety of interesting applications, including studies of chronic pain, responses to voices and distress, cardiovascular responsiveness to threat (Fielding-Smith et al., 2022; Peters et al., 2018), and the validity of retrospective reports of elderly people (Klumb & Baltes, 1999). In each context, the ESM provided useful information that might not have been possible to discover in a more formal laboratory setting. The ESM, especially if integrated with other data obtained under different methodological auspices, has the potential to expand our understanding of behavior.

Participatory Observation: When and Why

Given the problems of entry, of going native, and of observer interference in the natural actions of the group under investigation, to which studies employing observer participation are so susceptible, we might wonder why participatory observation would be used. Cicourel (1964) provided perhaps the best defense of this approach when he observed,

More intensive participation has the advantage of affording the observer a greater exposure to both the routine and the unusual activities of the group studied. The assumption is that the more intensive the participation on the part of the researcher, the 'richer' the data. (p. 44)

The personal relationships that can form in participant observation research can materially influence the quality of the data collected. For this reason, participatory observers are urged to be friendly, nonthreatening, concerned with the group's welfare, etc. Indeed, there is a growing literature in the social sciences demonstrating Guba's (1980, p. 21) dictum, "Good guys get better data." The observer can best demonstrate friendliness, humanness, etc., with the respondents when that researcher is in relatively frequent contact with the observed respondents in participatory observation research (Guba & Lincoln, 2008).

Reacting to the question of when participatory techniques should be used, Becker (1958) responded that sociologists usually use this method when they are especially interested in understanding a particular organization or substantive problem, but they assume that they do not know enough about the organization *a priori* to identify relevant problems and test hypotheses.

If we accept Cicourel's and Becker's delineations of the boundaries of this technique, then participatory research methods would seem most useful in exploratory, hypothesis-generating research, in which extensive "rich," if not necessarily reliable information is needed. The primary concern is the accumulation of large amounts of data, not on the data's reliability or validity.⁴ Once having identified a set of relevant behaviors, however, it follows that techniques involving far less observer participation should be favored. In this way, the potential interference that can be generated by the observer's presence in the group under investigation is circumvented. For the most part, the non-participatory methods are most deficient when studying issues which call for the greatest observer participation. Generally, non-participatory techniques require the observer conduct research in settings that restrict mobility of the observed group, thus enabling the observer to maintain isolation from the group and at the same time ensure the observed group remains within the observational setting. Restrictions of this type can be employed while simultaneously maintaining the naturalness of the observational settings, behaviors, and treatments.

At every point on the observer's participation to non-participation dimension, there are advantages to be gained, and issues to be avoided. The investigator must assess the relative costs of one method considering the advantages it offers. This assessment is dependent on many factors, most notably the amount of information one possesses about the issue under investigation, the availability of both settings and participants for observation, and the ethical propriety of the techniques of the chosen approach.

Coding Observations

Observational researchers must decide not only on the degree of participation in the observational study, but also how the actions and behaviors to be observed will be recorded and codified. Figure 13.1 presents a flowchart of the steps taken in developing a coding system to be used in observational research.

The Structured-Unstructured Dimension

Differences in theoretical orientations and research training backgrounds of social scientists have resulted in wide variations in the degree of structure of the coding systems they use in

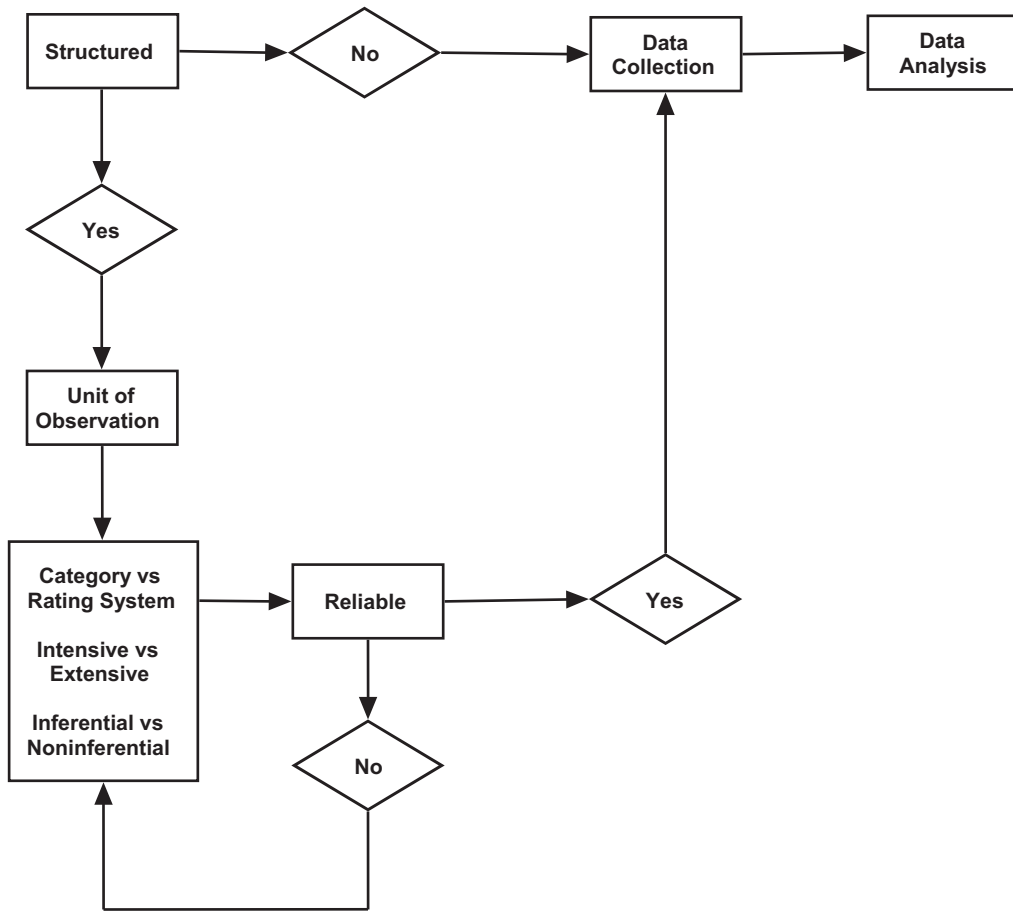


FIGURE 13.1 Decision tree used in developing a coding system.

observational research. At one extreme on this dimension are the proponents of complete systematization of procedures to collect and score observational data. These researchers argue that structure must be imposed on the mass of input accruing in any observational study to investigate hypothesized relationships. Common in research of this type are coding systems that focus on a limited and explicit portion of the possible range of behaviors that occur. In this way, reliable ratings of specific behaviors thought to have some theoretical importance can be isolated and tested. Researchers favoring less structured observational techniques (e.g., Glaser & Strauss, 1967) have argued against attempts to focus on a limited range of behaviors. Constriction of the observational field, they suggest, can result in a loss of valuable data that happen not to fall within the range of the chosen coding system. In unstructured approaches, researchers' observations precede the emergence of hypotheses or theory, based on interpreting observations as they unfold.

Early data returns. A potential difficulty with the unstructured approach to observational research is that conclusions based on initial observations inevitably influence interpretation of subsequent observations. In Rosenthal's (1966) volume on experimenter expectancy effects, the sequential analysis of "early data returns" represents a subtle but nonetheless real source of bias. Clearly, the most obvious way to avoid this potential source of bias is to reserve detailed inspection of results until all the data are collected.

Generalizability of observations. In addition to the potential confirmation bias resulting from the sequential analysis of data, the way unstructured observations are collected also poses validity problems. The question of whether the “discovered” relations are artifacts of the specific observational sample, or represent meaningful, generalizable discoveries is difficult to address in research of this type. When an observer has no predetermined categories of behavior to attend to, those events that stand out in the specific setting in which they were collected will be noticed (Borgida & Nisbett, 1977). These hyper-visible events, however, will not necessarily occur in different settings. Concentrating on the conspicuous occurrences of a situation can compromise the comparability of the obtained findings with those gathered in different settings. As Weick (1985, p. 44) has observed, “Routine activity is the rule, but it seldom gets recorded.” Bertrand Russell (1930) made essentially the same point when he observed that great people were great only infrequently – most of the time, their lives were relatively humdrum.

Consider the dilemma of observers who employ no prespecified observational coding system in their research. Literally hundreds of different behaviors occur in situations involving even very simple everyday activities. For example, suppose we were to observe the act of vacuuming – possible observable variables that constitute this simple chore range from the brand of vacuum used, size of vacuum, shagginess of the carpet, color of the carpet, the amount of time the participant vacuums, whether a primarily forward or backward motion was used in handling the contraption, amount of dirt collected in the dustbin, the socio-demographic characteristics of the person doing the vacuum, and so forth. Can the observer make any clear assertions regarding the generality of the behaviors that are observed if no *a priori* coding procedure was established? Probably not, since any obtained patterns might be specific to the sample of behaviors noticed in the early observations. The problem of the situation-specific determination of what is observed or noticed is avoided when the investigator enters the research field with variables representing predetermined categories of behavior to guide and structure the observational activities.

It would be unwise to interpret this critique of unstructured methods as a general condemnation of this approach. Insofar as the focus of these methods is generating testable hypotheses, we have no objection to their use. Barton and Lazarsfeld (1969, p. 182) advanced this position even more strongly, when they observed,

Research which has neither statistical weight nor experimental design, research based only on qualitative descriptions of a small number of cases, can nonetheless play the important role of suggesting possible relationships, causes, effects, and even dynamic processes. Indeed, it can be argued that only research which provides a wealth of miscellaneous, unplanned impressions and observations can play this role.

It is when the theories or hypotheses suggested by such methods are accepted, without first having undergone the necessary verification process in more controlled settings, that the potential dangers of unstructured observation become real. Unstructured techniques are valuable in generating theories, but their use in hypothesis-testing situations is always debatable.

Structured Methods: Category Systems

Whereas it might appear that unstructured techniques provide a simple means of conducting research, this is decidedly not the case. The ability to extract significant aspects of group or individual behavior, to filter from the mass of input those relations of potential theoretical importance, to reconstruct from necessarily incomplete notes the pattern and flow of the observational record is acquired only after intensive commitment to research of this type (Hoare et al.,

2012). It was partly for this reason, and partly in response to the call for more quantitatively based research techniques, that many social scientists reoriented their observational methods in the direction of greater structure. The most extreme point of this development is represented by research in which rating scales are used to quantify and summarize the actions of the person or group under study. Rather than begin at this extreme of structured observational techniques, we will direct our attention first toward methods that make use of category systems. These techniques fall somewhere near the midpoint of the structuredness dimension, between the completely open, unstructured varieties of research at one extreme, and the highly constricted rating scale approaches on the other.

Every category system represents an attempt to quantitatively summarize the qualitative behaviors that occur in the observational setting. In its most simplistic form, this approach might involve a simple count of the number of times specific types of events occur. A category is a specific description of a type of behavior, or other observable feature of a group or individual. A set of these descriptions, quite literally a checklist, constitutes a category system (or system of categories). Each of the categories represents the level of a variable being observed. Given a system of two or more categories, the observer's task is to note the number of times respondents' actions fall into one of the unique categories that constitute the variable. Categories of a variable should be mutually exclusive: an event coded as satisfying the demands for inclusion in a given category should not satisfy the demands of any other category. The question of the degree of intensity of occurrence is usually not a concern when category systems are used, unless different categories are defined by different intensities. Like a light switch (rather than a rheostat), the category is coded as either "on" or "off."

The unit of observation. A recurring issue in research making use of category systems concerns the unit of behavior to be considered when categorizing a social act – that is, when does one behavior stop and another begin? The problem of unit definition has plagued social scientists for some time, probably first inviting the attention of early quantitative anthropologists (e.g., see Murdock, 1967; Murdock & White, 1969). The question then, as now, concerns the problem of the definition of the appropriate unit of observation. The unit problem in observational research, though at a level of somewhat less grand proportion than that encountered by quantitative anthropologists, is troublesome nevertheless.

In many cases, solutions for specific category systems have been made by defining the observable units with respect to the categories being employed. For example, in his analysis of small group interaction, Bales (1950) chose as the observable unit of behavior "...the smallest discriminable segment of verbal or nonverbal behavior to which the observer...can assign a classification" (p. 37). In other words, any act that could be classified was counted as a unit of behavior.

Another possibility for defining the unit of observation is the use of a prespecified time interval. In Leventhal and Sharp's (1965) coding system of facial expressions, *for example*, during a given, prespecified time period, observers coded the expressions of the forehead and brow, then the eyes, then the mouth, etc. Given the physical restrictions of Leventhal and Sharp's observed group (women in labor) on whom this technique was employed, the time-determined unit offered a useful and unambiguous solution to the problem of the selection of the behavioral act for categorization. However, in observational studies of highly mobile individuals, such as young children at play, the definition of an action by temporal criteria could be inappropriate, because a target child could change behavior several times during a specified time interval. Accordingly, many observers of children's behavior have used the child's attention span in defining the unit of observation. So, if a child attends to a given object (be it another child, a teacher, a toy, etc.), this is a single unit of observation. When attention shifts to another object, the shift defines yet another observable "unit" for categorization. These examples represent the polar extremes of

unit determination, from the objectively defined unit, time, to the respondent-defined unit, attention shift. A combination of both approaches could be used; for example, another study may define an observable unit as children focusing attention on an object or activity for at least 15 seconds before it may be defined as a codable unit.

There are some coding systems in which the respondents themselves define the unit. For example, DePaulo and her colleagues (1996) were interested in studying lying in naturalistic settings. Information on many aspects of lying was collected, including the types of lies told, why, to whom, and so on. The researchers asked volunteers to keep a diary, and to note all of their daily social interactions that lasted more than ten minutes, and also all the lies they told during an interaction, no matter its duration. As the researchers were aware of the challenges in narrowly defining the features that precisely constituted a lie (e.g., whether a “little white lie” should be considered a lie), the participants were allowed to use their own personal judgment in defining the unit of lying – that is, they were to count as a lie any action they took intentionally that successfully misled another person. It is risky to allow participants to define the unit of analysis; some respondents, for example, might have minimized the reporting of their lying by justifying some forms of deception (e.g., lies that hurt another who “deserved” it) to not really be lies to appear more ethical. However, to adopt other unit-defining methods might have been difficult using the diary approach. Allowing the participant to define the unit was a conscious choice of the researchers that was dictated by the method.

There can be many gradations of unit definition, and no general rule to guide the researcher in choice of unit of observation other than common sense and the objectives of the research. As in DePaulo et al. (1996), it often is the case that the observational context itself will suggest, if not dictate, the unit and the way the unit is defined.

Time: When and how long. An important consideration in developing category systems concerns time. When using the temporal aspect of systematic observations, two general issues arise. The first of these has to do with the timing of observations. Some systems divide the behavioral stream into temporal units: over every time interval (say every five, ten, or 20 seconds), a behavioral code is assigned to the individual or group under observation. In this approach, time defines the unit. Research of this form often is used to study recurring interdependencies between behaviors. In a study of young children’s play behavior, for example, we might find that children’s aggressive outbursts are often preceded by other children frustrating their desires.

A second way in which time is used in observational research is not as a segmenting system, but as a dependent measure, signifying the duration of a particular coded event. Returning to our children’s example, we might code both the child’s aggressive outburst *and its duration*. This form of measurement is becoming increasingly popular with widely available video capabilities, which impart a running time signature on the interaction record (Bakeman, 2000).

Category System Construction

Given the number of category systems already available from previous research, how can the investigator choose one over the others? Or, having considered the possible choices and rejected them all, what are the criteria researchers should bear in mind in building their own systems? Here we consider some important aspects of category system construction.

Coding system intensity-extensity. An important question the prospective category system constructor must address concerns the degree of detail required of the observational data. Should the researcher pursue a general description of all broad events that occur in the observational setting or, rather, concentrate intensively on a particular set of precisely defined actions or behaviors, to the exclusion of all others? The answer depends on the topic being investigated.

Even with the same event, the level of observational scrutiny from intensive narrow detail (e.g., a smiling child with braces salivating over a cheeseburger) to broad extensive coverage (e.g., children eating lunch in a school cafeteria) might render a different picture.

Making use of an extensive system, which aims for broad coverage (at the expense of fine detail), an investigator can enhance the likelihood that all potentially relevant events occurring within the observational period will at least have been noted and categorized. It usually is the case, however, that the more extensive the category system, the less complete are the recorded (or coded) details of any given event made by the observers. Thus, whereas making use of an extensive coding scheme might show the general outlines of the total behavioral patterns, it is unlikely these data will convey much precise or nuanced information.

Increased precision of description demands constriction of observational focus. To describe an event intensively, the observer must ignore other incidents that happen to occur simultaneously. For example, if coders are instructed to pay attention exclusively to faces in categorizing emotional expression (e.g., joy, anger, fear, or sadness), observations of body movements in general (e.g., aggressive body stance) might be overlooked. In general, if one is concerned about a specific behavior or set of sequentially linked behaviors, the loss of broad, somewhat extraneous information (from the perspective of the coding scheme) might not be bothersome. However, valuable information may be lost when an overly intensive (and, thus, restrictive) observational scheme is adopted before other important features of the interaction are ignored.

There is an obvious trade-off involved in the choice of intensive versus extensive category systems. The decision regarding the type of system to use should be based primarily on the relative amount of information the investigator possesses about the phenomena of interest. In an exploratory study, intensive concentration on a specific set of behaviors (e.g., nonverbal displays) is risky because the arbitrarily chosen indicators might convey little worthwhile information. This might arise if all coded observations fall into the same category. If a researcher using a facial coding scheme has difficulty in discerning and therefore distinguishing among participants (as might be found in coding professional poker players in a high-stakes game), it is clear the researcher has not done the necessary homework to inform the choice of coding system. When the boundaries of the phenomenon in question have been clearly delineated and differences between categories can be observationally discerned by the coders, the use of an intensive system that focuses specifically on the events known to be important is indicated.

Most widely used category systems range between the extremes of complete coverage of broad events and intense concentration on minute behavioral details. A good example of a relatively extensive category system, which is limited to the classification of interpersonal behavior, is provided by Bales' (1950, 1970) interaction process analysis (IPA), presented in Figure 13.2. (See Bales and Cohen, 1979, and Nam et al., 2009, for an example of its application.) In this system, the 12 coding classifications represent a broad range of possible interactive behaviors that can occur between members of a small group.⁵

In contrast to Bales' extensive system, Caldwell's (1968) coding system, which she dubbed APPROACH (a procedure for patterning responses of adults and children), is a 65-category coding scheme that is highly restrictive in the events observed and recorded. In this case (as in most other systems of this type), an intensive categorization system reflects the researcher's desire to provide a system that is sensitive enough to allow for very fine distinctions to be made between highly similar behaviors.

Number of categories in the system. Apart from differing in terms of intensity/extensity of categories, the most striking contrast between Bales' and Caldwell's systems is the difference in the absolute number of categories used. The choice of number of categories for the variable being studied is an important consideration in constructing or choosing a coding system.

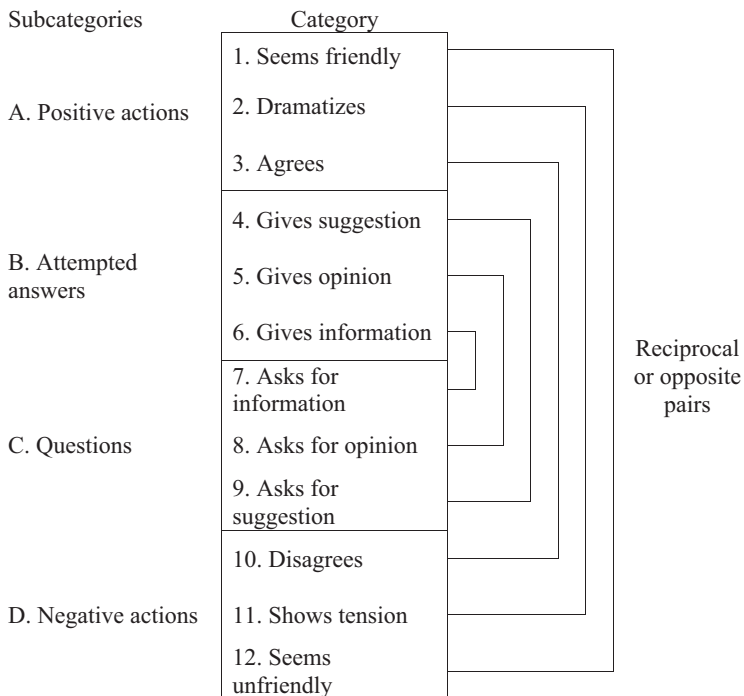


FIGURE 13.2 Categories of Bales' interaction process system: One variable with 12 categories.

Generally, the more complex the behavior to be observed, the simpler should be the category system. The logic of this assertion becomes clear upon examination. Keep in mind that with all coding systems, decisions on the part of coders regarding the assignment (i.e., categorization) of any given action are a vital part of the process. Further, in the study of any complex set of behaviors, even the most intricate of systems will not be so complete as to handle every action. Thus, the combination of highly complex observational data and many possible coding options (some of which require the observer to make extremely fine-grained judgments) can result in an overload for even the most experienced, well-trained coder.

Using fewer, broader categories will tend to transform the coders' function from one of split-second decision-making to that of more leisurely observation and categorization of events. In Figure 13.2, although 12 categories or levels may be coded, the researcher may choose to collapse and code into four categories instead (positive actions, attempted answers, questions, negative actions). Thus, instead of differentiating between the three types of positive actions (i.e., seems friendly, dramatizes, or agrees) as separate categories, such observations will be judged the same and simply be coded as positive actions. This results in a much easier coding job for the observers but will yield data insensitive to discriminating among the various types of positive actions.

With the widespread use of video recording, some of these issues are less important than has been the case in the past. Even so, despite the assistance of technology, it is still true that coders "burn out," so lessening the demands on them is a good long-term strategy. In much observational research, especially that conducted in settings that restrict respondents' movements, videos of the critical interactions can be made, and viewed as often as necessary to simplify reliable coding. In this way, complicated coding schemes can be used to study even complex behavior.

Such a combination will call for considerable re-reviews of the video, but with a sufficiently motivated cadre of coders, such a process can result in a highly informative analysis.

Studying less complex behavior demands a more elaborate classification system to ensure that the subtle behavioral differences that do occur will be noticed. An example of the type of behaviors that can be examined with a complex coding scheme is provided by Caldwell (1968), whose observers were to note the actions and reactions of individual children to their surroundings, a nursery school setting. In this setting, coders had ample time to determine which of the 65 APPROACH categories described the responses of the child they were observing. Using the same system to classify the responses of adults in a group discussion situation would have proved considerably more difficult, because the adult's actions would have been more subtle, exhibited greater purpose, and occurred with greater rapidity and frequency. In this instance, Bales' 12-category IPA would be a more realistic and appropriate option, unless the group interaction was videotaped.

Increasing the number of categorical variables. Thus far we have discussed only coding categories in single variable systems. It also is possible to categorize responses guided by multiple coding dimensions or variables. DePaulo et al. (1996) provide a good example of such an approach. In their research on lying, the researchers were interested in four different variables reflecting this behavior: the content of the lie, its reason, the type of lie, and the lie's referent. A schematic of this system is presented in Table 13.1. Researchers assigning categorical codes to the lies respondents mentioned in their diaries would code each lie according to each of the four variables of content, reason, type, and referent. Coding decisions for these four variable types provide more information and thus more variables for subsequent analyses. After data collection, two or more variables might be evaluated in statistical tests to determine if a relation or association exists between them. Results might show that lies typically classified as "feelings" on the content variable tended to also be classified as "protect liar" for the reason variable. The results would imply that lies to conceal actual feelings serve the social purpose of protecting the liar or the deceived individual from emotional harm.

Assessing category use. As was emphasized, constructing a useful coding system should involve careful consideration of the coding system, the observational context, and the types of behaviors to be observed. To construct a coding scheme independent of consideration of the settings in which it will be employed, and of the types of behaviors it will be used to classify is

TABLE 13.1 Coding scheme of Depaulo et al. (1996): four variables

<i>Variable</i>	<i>Category</i>
Content	1. Feelings 2. Accomplishments or knowledge 3. Actions or plans 4. Factors or property
Reason	1. Protect liar 2. Protect other
Type	1. Total falsehood 2. Exaggeration 3. Subtle (evasion or omission)
Referent	1. Liar himself/herself 2. Target of lie 3. Another person 4. Event or object

unwise. Thus, we recommend that the practicality and utility of any given coding system under construction be regularly and systematically investigated. Arbitrarily selecting a coding system often results in superfluous categories, or some classifications that should be broken down into smaller, more precise units of analysis. This is not meant as an argument to reinvent the wheel with each new study, but rather that the coding scheme should be compatible with the research purposes and terrain (see Bakeman & Casey, 1995, or Bakeman & Gottman, 1997, for more extended discussions of this issue).

In the initial phase of category construction, it is best to employ the planned system in a pilot test to note the frequency with which the various categories are used. In general, a good set of categories will be one in which all the category levels are used relatively frequently and evenly. If a given category is consistently under-used, it either should be broadened to account for more of the behaviors that occur or dropped as a category. Conversely, a category whose frequency of usage is considerably greater than all the others should be divided into more discriminating classifications. Category refinement of this nature should proceed until the distribution of descriptive classifications in the preliminary investigation is not extremely disproportionate. It is reasonable to use a category system when this phase of the construction process is completed.

Inference. In addition to the potential problems mentioned, the category system constructor is faced with a difficult decision when deciding upon the degree of inference the prospective system will require of coders. Should the classification system deal strictly with *observable events*, or, rather, require the coder to make some inferences regarding possible attitudes, motivations, or intentions that underlie the participant's actions? As psychological tendencies and states are not directly observable, information regarding these mental processes must be inferred from observable behaviors. If a coding variable is based solely on clearly observable behaviors, then the investigator need not be overly concerned about possible differences in interpretation of the same behavior by different observers. Such concrete, behavior-based category systems, however, often do not allow for the clear transmission of the "feeling" of the interaction. The subtle nuances of the observational situation are lost, often at cost to the interpretation and integration of the obtained results. Many years ago, Heyns and Zander (1953, p. 390) illustrated the choices confronting researchers:

At the simpler (noninferential) extreme, [the observer] might develop categories such as 'Shoves other children,' 'Calls other children names,' 'Asks for help,'....On the other hand, he may use such categories as 'Shows hostility,' 'Demands submission,' [etc.]. In the latter case the observers are looking at the same behavior but are making inferences concerning it.

When determining the inference level of categories, much depends on the degree to which the coding operations of the investigation are removed from the actual observational setting. If the research operations call for the concurrent observation and coding of behavior, then inferences regarding the motivation underlying the observed events can cause problems. Interpreting and coding a person's motives "online" is difficult. Coding online does not allow the coder the time to sit back and infer an actor's internal state. Inference-making is facilitated if the coding and categorization operations are video recorded and can be reviewed as frequently as necessary.

The predominant school of thought holds that observational methods be applied only to directly observable events and behaviors. Many value the methodology of systematic observation precisely because it is largely inference-free, although sometimes there are disagreements among raters observing the same event. Indeed, keeping coder judgments of underlying, unobserved cognitions to a minimum is bound to facilitate the reliability of a coding scheme. On the other hand, inferences about the motives, beliefs, or emotions underlying a given behavior can be

obvious sometimes. These inferential judgments add richness to the coding if they can be made reliably.

Interrater reliability. A primary consideration in the construction of any observation coding system concerns the issue of the reliability of observers in coding the categories of a variable. As Geller (1955, p. 194) noted, “The fewer the categories, the more precise their definition, and the less inference required in making classifications, the greater will be the reliability of the data.” Weick (1985, p. 38) outlined four types of comparisons that could be used in observational research to assess the extent of rating reliability of a coding variable. They involve the correlation between (1) ratings of two raters coding the same event, (2) ratings of a single rater coding the same event at different times, (3) ratings of two raters coding an event at two different times, and (4) the rating of a single rater watching a single event, but with ratings of the first half of the session correlated with those of the last half (a kind of split-half approach). The first of Weick’s comparison types is the most widely used when researchers report intercoder reliability.

In qualitative classification systems, nominal measurement scales are used, with values that have no meaningful quantitative differentiation in terms of whether a higher value reflects a higher score. Examples of qualitatively coded variables include participant biological sex (male/female), religion (Catholic, Protestant, Muslim, Buddhist, etc.), and whether the participant performed a particular action (yes or no). Another example of a qualitative rating system is the one presented in Table 13.1, which shows that for a particular observation, one of the four categorical levels (1 = feelings, 2 = accomplishments or knowledge, 3 = actions or plans, and 4 = facts or property) could be assigned for the content variable. The most fundamental question of interrater reliability is, “Do the ratings of two or more observers who have witnessed the same event(s) coincide to an acceptable degree?” By acceptable degree, we mean beyond that which would occur by chance. For example, if we have a system consisting of only two categories, and our two observers agree 60% of the time on this variable, should we be satisfied? Recognizing that agreement in such a system containing two categories would occur by chance 50% of the time, it is evident that a 60% intercoder agreement rate is not much to write home about.

Cohen’s kappa is an index to assess the extent of agreement between two or more coders of a qualitative categorical variable while controlling for chance (Cohen, 1960, 1968). Kappa’s value can range from .00 (no agreement whatsoever) to 1.0 (perfect agreement). Conceptually, the statistic is relatively straightforward. We begin with a matrix whose rows and columns reflect the various categories of the coding scheme. The rows represent the codings of Observer A, the columns those of Observer B. So for example, suppose we are using the category system that DePaulo et al. (1996) used to code lying, and are focusing on the first dimension of their system, namely the content of the lie. We could develop an agreement matrix by listing all the codings made by two observers for the content variable, as in Table 13.2. Each coder separately indicated the core content identified in each of the 236 lies. The diagonal entries in the table represent the number of times the coders have overlapped (i.e., agree) on a given category. The nondiagonal elements represent disagreements. The greater the number of entries off the diagonal, the lower the kappa and the lower the interrater reliability of the coding system.

To calculate Cohen’s kappa (κ), we first need information on the proportion of intercoder agreement. This is calculated by summing the diagonal entries and dividing by the total number of observations. In our example,

$$P_{\text{agree}} = (44 + 61 + 38 + 19) / 236, \text{ or } 162 / 236, \text{ or } .686.$$

The index reveals that 68.6% of the observations of content overlap between the two coders. To find the proportion of interrater agreements expected by chance, we multiply each column total

TABLE 13.2 Agreement matrix used in calculating Cohen's kappa: Two observers categorically coding 236 lies

		<i>Observer B</i>				
<i>Category</i>		<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>Total</i>
Observer A	1. Feelings	44	5	0	6	55
	2. Accomplishments	4	61	5	22	92
	3. Actions	2	5	38	0	45
	4. Facts	0	18	7	19	44
	Total	50	89	50	47	236

with its respective row total, and sum these products. This sum would then be divided by the square of the total number of observations.⁶

To determine Cohen's kappa, we use the following formula:

$$\kappa = \frac{P_{\text{agree}} - P_{\text{chance}}}{1 - P_{\text{chance}}}$$

Plugging our values into this formula, this produces $\kappa = .567$. The value is somewhat less than the proportion of intercoder agreement originally found ($P_{\text{agree}} = .686$), and reflects the adjustment made after ruling out chance agreements.⁷ This is not a particularly strong result. Generally, observational researchers suggest that a kappa greater than .75 is "excellent"; kappas between .60 and .75 are "good," .40 to .59 is "fair," and those less than .40 indicate a "poor" intercoder agreement (Fleiss, 1981).

The concordance matrix of Table 13.2 gives some indication of the source of the relatively low kappa and, by extension, its solution. To improve the coding scheme, or the coder training, we scrutinize the off-diagonal entries. Are there cells off the diagonal that suggest an inordinate or unacceptable number of disagreements? In the hypothetical table presented here, our two observers had most disagreement on the categorization of lies as involving either accomplishments (Category 2) or facts (Category 4), as these reveal the highest number of non-correspondences. This suggests that the researcher must clarify the operational definition of these two categories to the coders, combine or remove categories, or spend more time training coders about the categorical distinctions – or all three.

Fleiss' kappa (1971) is a more general index to assess the extent of agreement between two or more coders of a qualitative categorical variable while controlling for chance. Fleiss' kappa ranges on a reliability index from 0 to 1.0, with the same cutoff criteria involving "excellent," "good," "fair," and "poor" for Cohen's kappa. With exactly two coders, it is acceptable to use Fleiss' kappa to calculate interrater reliability, as the resultant value will be the same as Cohen's kappa (except for rounding error).

Weick (1985) holds that the reliability of a coding system reflects both the discriminability of the classifications that constitute the scheme and the efficacy of coder training. In the construction of category systems, interrater reliability is of major importance, because without it very little use can be made of the collected observations. Enhancing coder agreement often calls for more intensive coder training. Given sufficient practice with a coding scheme, most coders can learn to employ even the most complex systems reliably.

Nondiscriminability of categories, however, represents a more difficult problem. If a single unit of behavior can be logically classified using two or more categories contained in the same

coding variable, it gives rise to intercoder disagreements that augmented training will not resolve.⁸ The appropriate response to problems of this type is to restructure categories to delineate more clearly the boundaries of each, by combining categories posing challenges in coder discrimination. This, in turn, calls for additional coder training to determine whether the sought-for effectiveness in category differentiation has occurred.

All the previously mentioned aspects of system construction affect the reliability of the coding scheme. To summarize, the interrater reliability of a coding variable will be enhanced if:

- The coding variable is extensive, rather than intensive.
- The unit of analysis is predefined by the observer, rather than based on some dynamic action of the observed.
- The coding variable consists of a small rather than large number of categories.
- Actions of the observed and classification by observers are not concurrent processes (i.e., audio- or video devices are used to allow for careful review of the observational data).
- Only directly observable behaviors are assessed, so that little or no psychological inference is required.

We do not mean to suggest that these considerations should always serve as the central criteria to determine how categories are constructed for each variable, but rather to call attention to the fact that some types of categorization are less susceptible to problems of interrater unreliability.

Structured Methods: Rating Scales

Sometimes the research is not concerned with classifying observed events into distinct qualitative categories, but rather quantitatively rating observations in terms of frequency, amount, or duration. In its simplest form, a quantitative rating scale is an extension of a qualitative system, in which a coder classifies by evaluating the magnitude or intensity of a variable. Quantitative observations may be ordinal, interval, or ratio (Hallgren, 2012), with higher values representing greater quantities of the observed variable. Examples of quantitative coding variables include the amount of time (in seconds or minutes) that a lie was elaborated on. Another example is the frequency of times the same lie is told. To make inferences about non-directly observable psychological or attitudinal events, a coding variable might use quantitative ratings with endpoints to help guide judgments. For example, coders might be required to evaluate the persuasiveness of each lie, from 1 (very unconvincing) to 10 (very convincing). This additional aspect of coding often makes achieving interrater reliability more difficult and alters the way it is assessed.

The **intraclass correlation** (ICC) is an index of interrater reliability for two or more coders of a quantitative variable (Shrout & Fleiss, 1979). The ICC ranges from 0 to 1.0, with higher values representing stronger interrater reliability for that variable. Recall from Chapter 11 that this is the same ICC that is used in analyses of multilevel modeling and has a similar logic: Observations are nested under each coder. In other words, each coder rates observations for all participants in a study. In this case, the ICC refers to the degree to which two or more coders' observations overlap.

Suppose that each of four observers independently provided a rating, based on the previously mentioned 10-point persuasiveness scale, for 15 lies. The judgments are shown in Table 13.3. The ICC would be used to provide a summary of the extent that observers' ratings are related. Unlike Cohen's kappa and Fleiss' kappa, both of which require an exact correspondence in categorization among coders to be deemed a match, the ICC does not demand identical scores from coders to yield a relatively acceptable index. Whereas qualitative rating scales are concerned

TABLE 13.3 Interrater scores used in calculating the ICC: Four observers quantitatively rating 15 lies

Lie #	Observer A	Observer B	Observer C	Observer D
1	9	9	10	10
2	6	5	6	5
3	2	4	5	9
4	2	4	3	4
5	3	2	5	3
6	6	5	7	7
7	5	6	8	6
8	2	5	4	6
9	7	7	7	7
10	3	3	5	2
11	8	9	1	2
12	3	3	5	2
13	9	10	7	8
14	7	9	7	4
15	5	5	5	6

with *absolute agreement* among raters (coded categories do or do not match), quantitative scales are more concerned with consistency or *relative agreement* among raters (the extent of the match among raters). A high ICC will be achieved if the pattern of rated scores assigned by a coder is relatively consistent and in the same direction as the scores for the other coder(s). Thus, if one coder's mean scores are consistently higher than another's, but their patterns of scores are highly similar, a strong ICC will be found.

Using the values shown in Table 13.3, the calculation of the ICC yields a respectable interrater reliability of .80. Although the statistical details are beyond the scope of the chapter, the ICC may be understood as the extent of correlation among the scores made by multiple raters. Let's try to understand how certain patterns of obtained scores may impact the value of the ICC. Inspecting Table 13.3, notice that for lie #1, the four coders did not assign precisely the same values. Although not identical, the ratings assigned for this lie will tend to contribute to a higher ICC, because the relative agreement pattern shows that all four raters typically assigned higher scores. Conversely, lie #3 would contribute to a lower ICC, because the categorizations across all four judges were erratic: Observer 1 gave a low score, observers 2 and 3 gave middling scores, and observer 4 gave a high score. In fact, if this lie had been omitted from the analysis, the remaining 14 lies would yield a higher recomputed ICC of .84.

Researchers should understand the circumstances in which it is suitable to use the ICC versus Cohen's kappa versus Fleiss' kappa in assessing interrater reliability. If the quantitative ratings from Table 12.3 were accidentally used to compute Fleiss' kappa, this would yield an assessment of absolute coder correspondence for the ten categories that would be erroneously low and therefore a meaningless interrater reliability value.

Combining Quantitative and Qualitative Coding

Different qualitative and quantitative coding variables may be integrated into the same observational study. Suppose a researcher coded observations based on the qualitative variable of *reason for lying* (protecting the liar or protecting the other person) and the quantitative variable of persuasiveness (using a 1- to 10-point scale) of that same lie. Suppose that this system was used

by observers to code the lies, and both variables were found to yield strong interrater reliabilities. After determining acceptable systematic agreement among raters, these results allow the investigator to go forward to use these reliable variables in statistical analyses. A t-test analysis might show that the persuasiveness score averaged 8.5 if the reason for the lie is to protect the liar, but a significantly lower average persuasiveness score of 4.0 is found if the reason concerned protecting others.

Consider the following problem: An investigator uses Bales' IPA to code the behavior of an individual in a small group interaction. The participant's response to another person in the group must be classified using one of the 12 IPA categories. Suppose further, the observer must determine qualitatively the proper classification of an action and estimate quantitatively the intensity of the response. Having decided that an act by person A "seems friendly" (Category 1) toward person B, the coder must then quantify whether this friendliness was 1 = *very low*, 2 = *somewhat low*, 3 = *average*, 4 = *somewhat high*, or 5 = *very high*. The result of this combined approach is that the complexity of the coders' task is greatly increased because they must not only categorize, but also rate the critical behaviors in terms of intensity, magnitude, duration, etc.

In opposition to these disadvantages stands the possibility of increasing the information value of the obtained results. Returning to Bales' IPA, suppose that after you've mastered observational coding, you travel back in time to the early days of the founding of the United States. You find yourself observing the members of the Second Continental Congress in 1776 during their deliberations on the Declaration of Independence. One member of the Congress, feeling great ambivalence over Jefferson's document, hesitantly states, "Yes, I suppose I can go along with this." You code this as an instance of Category 3, "agrees." Another member, perhaps better understanding the historical significance of his actions, jumps up and screams, "I will stake my fortune and my life on this document!" This action also is coded as an instance of Category 3. Clearly, these two acts are different, yet both receive the same score – the categorical (but intensity-free) IPA cannot denote even this obvious distinction. Through the inclusion of quantitative coding, such variable indications of the strength of the agreement from 1 (some agreement) to 5 (full agreement) are not glossed over, and a more descriptive picture of the social interaction emerges.

Conclusion

In observational research, the degree of control placed on the allowable actions of the observer is wide ranging. At the least controlled pole, we have the open, unconstrained methods of the early ethnographers, whose approach to the information accumulation is different from that of the highly structured, often ritualistic practices of today's experimentalists. We believe the development of increasingly sophisticated methods to reduce the impact of the observer on the observed reflects the growth, evolution, and overall progress of the science of human behavior. Yet, as we suggested at the beginning of this chapter, it often is the case that the more elementary, basic, "foundation building" observational techniques have been neglected in favor of the more "advanced" experimental ones. This can be a mistake. Schwartz and Jacobs (1979, p. 327) highlighted the importance of observational methods when they stated,

Ordinarily (in the experimental methods), one has a hypothesis, knows pretty well what it means, and wants to know if it is true. Here (in the observational methods), one has a hypothesis, is pretty sure that is true, and wants to know what it means (wants to know what is true.).

Notes

- 1 In Moriarty's (1975) study, 95% of those asked to watch the "victim's" belongings did intervene in the theft, whereas in a similar study that did not include a specific request, no one even tried to stop the staged theft of an expensive calculator in a college library (Austin, 1979).
- 2 Whether a study of this kind would be viewed today as ethically defensible is an open question. We discuss questions of this type in Chapter 20.
- 3 The basic ESM study requires a two-level nested design: Measurements of experiences at the different time points are nested within each participant.
- 4 In fairness, it should be noted that Becker has found the participatory observational approach suitable for hypothesis- and theory-testing, but he indicated that this operation could legitimately occur only after a number of prior requirements were satisfied (also see Kidder, 1981).
- 5 Note, however, Bales' categories deal with the nature of the interaction, not its content.
- 6 In our example, this process would require the following calculations:

$$P_{\text{chance}} = [(50 \times 55) + (89 \times 92) + (50 \times 45) + (47 \times 44)] / 236^2 = [(2750) + (8188) + (2250) + (2068)] / 55696.$$

The result of this series of calculations is 15256/55696, or .274 (27.4% of observations are due to chance agreement).

- 7 A number of online programs are available to reduce the drudgery of calculating kappa (e.g., <http://vassarstats.net/kappa.html>).
- 8 A good indication of nondiscriminability of categories occurs when the same observer, reviewing the same behaviors (typically through the use of video recordings), assigns different scores from the first to the second observation. Results of this type suggest that the coding classifications cannot be employed with any degree of reliability.

References

- Atkinson, P., & Hammersley, M. (1994). Ethnography and participant observation. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 248–261). Sage.
- Austin, W. G. (1979). Sex differences in bystander intervention in a theft. *Journal of Personality and Social Psychology*, 37, 2110–2120.
- Bakeman, R. (2000). Behavioral observation and coding. In H. Reis & C. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 138–159). Cambridge University Press.
- Bakeman, R., & Casey, R. L. (1995). Analyzing family interaction: Taking time into account. *Journal of Family Psychology*, 9, 131–143.
- Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). Cambridge University Press.
- Bales, R. F. (1950). *Interaction process analysis*. Addison-Wesley.
- Bales, R. F. (1970). *Personality and interpersonal behavior*. Holt, Rinehart, and Winston.
- Bales, R. F., & Cohen, S. P. (1979). *SYMLOG: A system for multiple level observation of groups*. Free Press.
- Barton, A. H., & Lazarsfeld, P. H. (1969). Some functions of qualitative analysis in social research. In G. J. McCall & J. L. Simmonds (Eds.), *Issues in participant observation*. Addison-Wesley.
- Becker, H. S. (1958). Problems of inference and proof in participant observation. *American Sociological Review*, 23, 652–660.
- Bogden, R. (1972). *Participant observation in organizational settings*. Syracuse University Press.
- Borgida, E., & Nisbett, R. F. (1977). The differential impact of abstract versus concrete information on decisions. *Journal of Applied Social Psychology*, 7, 258–271.
- Bourgeault, I. L. (2000). Delivering the "new" Canadian midwifery: The impact on midwifery of integration into the Ontario health care system. *Sociology of Health and Illness*, 22, 172–196.
- Caldwell, B. M. (1968). A new "approach" to behavioral ecology. In J. P. Hill (Ed.), *Minnesota symposia on child psychology* (Vol. 2, pp. 74–108). University of Minnesota Press.
- Carlson, J., Cook, S. W., & Stromberg, E. L. (1936). Sex differences in conversation. *Journal of Applied Psychology*, 20, 727–735.
- Cicourel, A. V. (1964). *Method and measurement in sociology*. Free Press of Glencoe.
- Cohen, J. A. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.

- Cohen, J. A. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70, 213–220.
- DeLongis, A., Hemphill, K. J., & Lehman, D. R. (1992). A structured diary methodology for the study of daily events. In F. B. Bryant, J. Edwards, R. S. Tindale, E. J. Posavac, L. Heath, E. Henderson & Y. Suarez-Balcazar (Eds.), *Methodological issues in applied social psychology*. (pp. 83–109). Plenum Press.
- DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., & Epstein, J. A. (1996). Lying in everyday life. *Journal of Personality and Social Psychology*, 70, 979–995.
- DeWalt, K. M., & DeWalt, B. R. (2010). *Participant observation: A guide for fieldworkers* (2nd ed.). AltaMira Press.
- Festinger, L., Riecken, H. W., & Schachter, S. S. (1956). *When prophecy fails*. University of Minnesota Press.
- Fielding-Smith, S. F., Greenwood, K. E., Wichers, M., Peters, E., & Hayward, M. (2022). Associations between responses to voices, distress and appraisals during daily life: An ecological validation of the cognitive behavioural model. *Psychological Medicine*, 52(3), 538–547.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76, 378–382.
- Fleiss, J. L. (1981). *The design and analysis of clinical experiments*. Wiley.
- Gable, S. L., Reis, H. T., & Elliot, A. J. (2000). Behavioral activation and inhibition in everyday life. *Journal of Personality and Social Psychology*, 78, 1135–1149.
- Geller, E. (1955). Systematic observation: A method in child study. *Harvard Educational Review*, 25, 179–195.
- Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Aldine.
- Guba, E. (1980). The evaluator as instrument. Unpublished ms., Indiana University.
- Guba, E. G., & Lincoln, Y. S. (2008). Paradigmatic controversies, contradictions, and emerging confluences. In N. K. Denzin & Y. S. Lincoln (Eds.), *The landscape of qualitative research* (3rd ed., pp. 255–286). Sage.
- Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, 8, 23–34.
- Henle, M., & Hubbell, M. B. (1938). ‘Egocentricity’ in adult conversation. *Journal of Social Psychology*, 9, 227–234.
- Heyns, R. W., & Zander, A. F. (1953). Observation of group behavior. In L. Festinger & D. Katz (Eds.), *Research methods in the behavioral sciences*. Holt, Rinehart, and Winston.
- Hoare, K. J., Mills, J., & Francis, K. (2012). Sifting, sorting and saturating data in a grounded theory study of information use by practice nurses: A worked example. *International Journal of Nursing Practice*, 18, 582–588.
- Howard, W. D., & Crano, W. D. (1974). Effects of sex, conversation, location, and size of observer group on bystander intervention in a high risk situation. *Sociometry*, 66, 255–261.
- Itzhakov, G., & Reis, H. T. (2021). Perceived responsiveness increases tolerance of attitude ambivalence and enhances intentions to behave in an open-minded manner. *Personality and Social Psychology Bulletin*, 47(3), 468–485.
- Kidder, L. H. (1981). Qualitative research and quasi-experimental frameworks. In M. B. Brewer & B. E. Collins (Eds.), *Scientific inquiry and the social sciences* (pp. 226–256). Jossey-Bass.
- Klumb, P. L., & Baltes, M. M. (1999). Validity of retrospective time-use reports in old age. *Applied Cognitive Psychology*, 13, 527–539.
- Kubey, R., Larson, R., & Csikszentmihalyi, M. (1996). Experience sampling method applications to communication research questions. *Journal of Communication*, 46, 99–120.
- Larson, R., & Csikszentmihalyi, M. (1978). Experiential correlates of solitude in adolescence. *Journal of Personality*, 46, 677–693.
- Larson, R., & Csikszentmihalyi, M. (1980). The significance of time alone in adolescent development. *Journal of Current Adolescent Medicine*, 2, 33–40.
- Larson, R., & Csikszentmihalyi, M. (1983). The experience sampling method. In H. T. Reis (Ed.), *Naturalistic approaches to studying social interaction* (pp. 41–56). Jossey-Bass.
- Larson, R., Csikszentmihalyi, M., & Graef, R. (1980). Mood variability and the psychosocial adjustment of adolescents. *Journal of Youth and Adolescence*, 9, 469–490.

- Leigh, B. C. (1993). Alcohol consumption and sexual activity as reported with a diary technique. *Journal of Abnormal Psychology, 102*, 490–493.
- Leventhal, H., & Sharp, E. (1965). Facial expressions as indicators of distress. In S. Tomkins & C. E. Izard (Eds.), *Affect, cognition, and personality* (pp. 296–318). Springer.
- Iida, M., Shrout, P. E., Laurenceau, J.-P., & Bolger, N. (2012). Using diary methods in psychological research. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics*. (pp. 277–305). Washington, DC: American Psychological Association.
- Mehan, H., & Wood, H. (1975). *The reality of ethnomethodology*. Wiley.
- Moriarty, T. (1975). Crime, commitment, and the responsive bystander: Two field experiments. *Journal of Personality and Social Psychology, 31*, 370–376.
- Murdock, G. P. (1967). Ethnographic atlas: A summary. *Ethnology, 6*, 109–236.
- Murdock, G. P., & White, D. R. (1969). Standard cross-cultural sample. *Ethnology, 8*, 329–369.
- Nam, C. S., Lyons, J. B., Hwang, H.-S., & Kim, S. (2009). The process of team communication in multi-cultural contexts: An empirical study using Bales' interaction process analysis (IPA). *International Journal of Industrial Ergonomics, 39*, 771–782.
- Nezlek, J. B., Wheeler, L., & Reis, H. T. (1983). Studies of social participation. In H. T. Reis (Ed.), *Naturalistic approaches to studying social interaction* (pp. 57–75). Jossey-Bass.
- Østerlund, C., Snyder, J., Sawyer, S., Sharma, S., & Willis, M. (2016). Documenting work: From participant observation to participant tracing. In K. D. Elsbach & R. M. Kramer (Eds.), *Handbook of qualitative organizational research: Innovative pathways and methods* (pp. 391–400). Routledge/Taylor & Francis Group.
- Peters, B. J., Reis, H. T., & Jamieson, J. P. (2018). Cardiovascular indexes of threat impair responsiveness in situations of conflicting interests. *International Journal of Psychophysiology, 123*, 1–7.
- Polifroni, M. A., von Hippel, W., & Brewer, M. B. (2001). Nazis on the Web: Support for image theory in the internet communications of white supremists. Paper presented at the annual meeting of the Midwestern Psychological Association, Chicago, IL (May).
- Przemieniecki, C. J., Compitello, S., & Lindquist, J. D. (2020). Juggalos—Whoop! Whoop! A family or a gang? A participant-observation study on an FBI defined 'hybrid' gang. *Deviant Behavior, 41*(8), 977–990.
- Ratliff, E. A. (1999). Women as “sex workers,” men as “boyfriends”: Shifting identities in Philippine go-go bars and their significance in STD/AIDS control. *Anthropology and Medicine, 6*, 79–101.
- Reis, H. T., & Gable, S. L. (2000). Event-sampling and other methods for studying everyday experience. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 190–222). Cambridge University Press.
- Roberts, R. R., & Renzaglia, G. A. (1965). The influence of tape recording on counseling. *Journal of Consulting Psychology, 12*, 10–16.
- Robinson, J. P., & Goodbey, G. (1997). *Time for life: The surprising ways Americans use their time*. Pennsylvania State University Press.
- Rosenthal, R. (1966). *Experimenter effects in behavioral research*. Appleton-Century-Crofts.
- Russell, B. (1930). *The conquest of happiness*. Horace Liveright.
- Schwartz, H., & Jacobs, J. (1979). *Qualitative sociology: A method to the madness*. Free Press.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428.
- Silvia, P. J., & Cotter, K. N. (2021). *Researching daily life: A guide to experience sampling and daily diary methods*. American Psychological Association.
- Thomann, S., Zwakhalen, S., Siegrist-Dreier, S., & Hahn, S. (2022). Restraint practice in the somatic acute care hospital: A participant observation study. *Journal of Clinical Nursing*, Advanced online publication. <https://doi.org/10.1111/jocn.16322>
- Thorne, B. (1988). Political activist as participant observer: Conflicts of commitment in a study of the draft resistance movement of the 1960s. In P. C. Higgins & J. M. Johnson (Eds.), *Personal sociology* (pp. 133–152). Praeger.
- Tunnell, G. B. (1977). Three dimensions of naturalness: An expanded definition of field research. *Psychological Bulletin, 84*, 426–437.

- Watt, S., & Jones, J. S. (2010). Let's look inside: Doing participant observation. In J. Scott Jones & S. Watt (Eds.), *Ethnography in social science practice*. (pp. 107–125). Routledge.
- Weick, K. E. (1985). Systematic observational methods. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (3rd ed., pp. 567–634). Addison-Wesley.
- Williamson, J., & Karp, D. (1977). *The research craft: An introduction to social science methods*. Little-Brown.
- Zegiob, L. E., Arnold, S., & Forehand, R. (1975). An examination of observer effects in parent-child interactions. *Child Development*, 46, 509–512.

14

CONTENT ANALYSIS

Content analysis refers to a research method for classifying, evaluating, and studying recorded communications in a systematic and objective manner. Information, messages, and content are embedded in recorded communications that offer valuable research insights about culture, social norms and values, and human activity. Our human ancestors communicated through recorded mediums such as cave drawings, carvings on pottery and ceramics, and other ancient artifacts. These antiquities have been studied by modern researchers using content analysis approaches. Today, people communicate and exchange informational content through modern recorded mediums such as newspapers, photographs, videos, and the Internet. All these sources have been subjected to content analysis research. Information from media sources – images, photographs, newspapers, audio clips, song lyrics, television shows, billboard advertisements, memes, tweets, online message forums, social media profiles, and websites – can be examined using content analysis (Lac, 2016; Stemler, 2001; Neuendorf, 2011). Additionally, raw unstructured information collected through research methods such as observational information, interviews, and focus groups can be used in content analysis to make sense of the recorded data.

Essentially, content analysis is a research approach for sorting, interpreting, and understanding unstructured content conveyed through communication mediums.¹ Keep in mind that “communication” is broadly defined and is not limited to traditional text-based communications. A wide variety of research questions concerning content found in various media and multimedia communications can be evaluated using content analysis. These include paintings (“Are children more likely to be depicted in competitive or cooperative play in Normal Rockwell paintings?”), movies (“Do American action movies usually portray Muslim characters as heroes or villains?”), billboards (“How much sexualization is depicted in car advertisements that target men compared to women?”), comic strips (“Do comic strips about the current President usually apply sarcasm, satire, or bathroom humor?”), and social media (“Does an association exist between the attractiveness of a person’s profile picture and how many friends that person has on Facebook?”).

Content Analysis Basics

As with observational coding systems discussed in Chapter 13, content analysis is broadly classified in terms of whether the developed coding system (codebook) that will be used to code

the content of interest will be represented as qualitative or quantitative variables. An example involving the content analysis of *memes* will help to understand this distinction. A meme is a digital object with text overlaid on an image background (picture or drawing) intended to be disseminated via the Internet and social media platforms (Shifman, 2013). Several published studies have performed content analyses of memes, including memes disseminated by political parties (Martínez-Rolán & Piñeiro-Otero, 2016) and a religious organization (Brubaker et al., 2017).

In a *qualitative content analysis*, the researcher reviews open-ended unstructured content by identifying themes, characteristics, and patterns using only descriptive words, sentences, and quotes. If performing a content analysis of memes about methamphetamine use, a researcher may review the type of negative consequence depicted in every meme. Afterward, the researcher may qualitatively summarize that information as follows: “The types of consequences commonly depicted in memes about methamphetamine use included decaying teeth, skin lesions, and premature wrinkles.” For a qualitative variable about humor, the researcher may write that “The memes typically mocked the methamphetamine users in a sarcastic manner.”

In a *quantitative content analysis*, the open-ended unstructured content is rated and coded into quantitative variables that can later be entered into statistical analyses. If examining memes about methamphetamine use, quantitative variables might assess whether each meme depicted negative consequences using variables of decaying teeth (0 = no, 1 = yes), skin lesions (0 = no, 1 = yes), premature wrinkles (0 = no, 1 = yes), in addition to humorous tone (1 = serious to 5 = humorous). Afterward, analyses could be performed to determine whether these quantitative variables are statistically associated. For example, an independent t-test may be conducted to determine whether memes depicted decaying teeth versus those that do not significantly differ in ratings of humorous tone. The major distinction is that words and phrases represent the coded variables in a qualitative content analysis, whereas numerical values represent the coded variables in a quantitative content analysis. These two approaches are not necessarily mutually exclusive. A *mixed-methods content analysis* involves some qualitative variables that are coded using words and other quantitative variables that are coded using numbers in the same study. This chapter focuses on quantitative content analysis, but many of the concepts presented are applicable to understanding qualitative content analysis.

Another decision that must be made is whether to specify the procedures used in the coding system ahead of time or devise them during the coding process (Kondracki et al., 2002). In **deductive content analysis**, the researcher determines the variables, operational definitions, and response options that will be used for the coding system prior to conducting the coding. This theory-driven strategy is based on hypotheses and models drawn from previous research on the topic, which specify the variables that are important to cull from the source data and the purpose of the research. In **inductive content analysis**, the researcher allows variables, operational definitions, and response options to emerge by continually revising and updating the coding system during the coding phase. This exploratory data-driven strategy might be used if little existing research knowledge exists about the topic, so the researcher is required simultaneously to develop the coding scheme while coding. Typically, inductive content analyses should be avoided, but if used, it is probably most appropriate during the pilot study stage to inform the creation of variables and refine operational definitions to develop the final codebook that will be applied to the final sample of content units. It is most meaningful to develop a systematic coding scheme prior to examining the vast amount of unstructured content.

Conducting a Content Analysis

Before describing the ways that content analyses are undertaken, we should consider some necessary considerations. Recognizing the similarity between the research procedures used by the

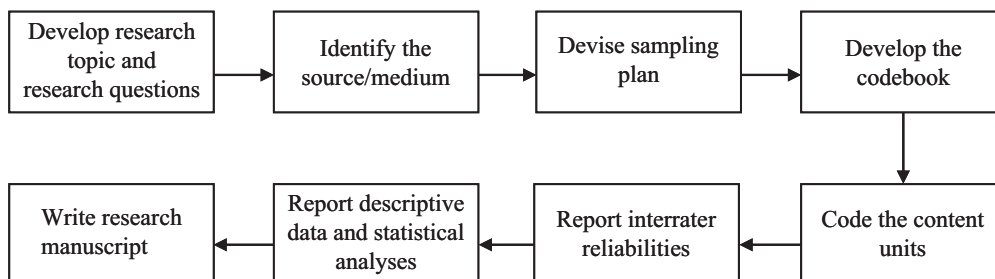


FIGURE 14.1 Steps for developing a content analysis research project.

content analyst and those of observational methods described in the last chapter will facilitate understanding the information in this chapter. Figure 14.1 provides a roadmap of the decisions that must be made in the process of mounting a content analysis.

Identifying the Source

Before beginning a content analysis, the investigator must first determine whether the technique is compatible with the content source and research goals. The content must be previously recorded or archived in some way for researchers to perform a content analysis. But not all sources for the content analysis may be accessible to a researcher, especially if the content is not available on the Internet, found in some library archive, or publicly accessible. Recorded sources and mediums of communication for a content analysis may be classified into four types: textual or written (e.g., newspaper articles, books, books, Reddit posts), auditory (e.g., music, speeches, radio broadcasts), visual (e.g., images, paintings, photographs), and multimedia (contains a combination of textual, auditory, and visual).

Another imperative issue that should be pondered is whether a source or medium is available that may be culled to provide the data necessary to answer the research questions. The decision about the selected source for the content analysis should be matched up to the research questions. If a researcher hopes to conduct a content analysis of popular children's songs from 1,000 years ago, such music or lyrics must have been previously recorded in some medium, otherwise that information is lost in time. If interested in the topic of methamphetamines, various recorded content sources and mediums are possibilities for the study. A content analysis on the topic of methamphetamines in the media can be conducted using one of the following sources: newspaper articles, newspaper cartoons, websites, anti-drug billboards, online message forums, memes, Twitter tweets, Instagram posts, or Facebook pages devoted to methamphetamines.

Creating the Codebook

After deciding on the source, the next step is to develop the codebook that represents the variables to be coded in the content analysis. Each coded variable should possess three characteristics: variable name (e.g., humorous), operational definition (e.g., "Rate the extent that the meme is intended to be humorous compared to serious. Humorous is defined as intending to be funny, silly, amusing, comic, absurd, ironic, or sarcastic. Judge based on the intention of the meme not whether you personally think is humorous"), and the response options (e.g., 1 = serious to 5 = humorous). It is important that the operational definitions and response options provide the sensitivity necessary to address the research questions and hypotheses.

Moreover, each coded variable in a content analysis codebook may be considered a manifest or latent variable. A **manifest variable** is a directly observable characteristic of the content and therefore requires little to no inferences to be made by the coder. On the other hand, a **latent variable** involves a characteristic of the content that is not directly observable and therefore requires inferences to be made by the coder. This concept is probably best viewed on a continuum from manifest to latent, with some variables possibly possessing manifest and latent features.

Continuing with the methamphetamine memes example, whether the consequence of decaying teeth is shown in the image is considered an example of a manifest variable. The coder could inspect the meme and directly observe whether the picture of the person depicted in the meme shows poor dental hygiene. The extent that each meme is humorous is an example of a latent variable, as the coder needs to judge the extent the image and text depicted in the meme are intended to be funny. Inferences might be required because what is judged to be highly humorous to one person might not be as humorous to another. Some researchers prefer not to make any inferences – only directly observed events (or behaviors) are coded in the form of manifest variables. Other researchers may make inferences about the motivations or intentions that appear to underlie the observed content in the form of latent variables. Inferences are thought to provide a richer and more meaningful analysis of the topic under study. Most published content analyses involve a mixture of manifest and latent variables.

Pilot Study

After devising the coding system of variables for the codebook, it is wise to test it on a pilot sample of the content units. The coders practice and are trained on this pilot content. Researchers conduct “pilot studies” of this type to answer a simple question: Do the coding variables exhibit acceptable interrater reliabilities? If not, modify the coding system based on the feedback and preliminary results from the coders. If, after modification, coders can agree upon codes they have applied independently to content, the researcher is ready to move on to the formal coding phase that involves having the coders rate the final set of sampling units. In this next step, the content units of interest for the entire sample are gathered. The coding may be done manually or by computer, but even computerized coding methods involve considerable human guidance. The data extracted from this process represent variables to be used in subsequent statistical analyses and interpretations.

Developing the Coding Variables

To understand how content analysis is conducted, it is critical to define the terms coding unit, coding variable, and context unit. A **coding unit** represents the fundamental unit of interest that should be attended to when performing a content analysis. For example, in a content analysis of obituaries of people who have died from COVID-19, a researcher may search for the coding unit represented by the word “pandemic.” The **context unit** is the embedded and wider context within which the meaning of the coding unit is to be interpreted by the researcher. For example, this may require reading the entire obituary to deeply understand how the word “pandemic” was framed. A **coding variable** represents the presence, frequency, or quantity of that coding unit to be recorded into a dataset and later used as variables in statistical analyses. For example, the number of times that the word “pandemic” is mentioned in each obituary represents the coding variable.

Sometimes the coding unit and context unit are identical, but more likely the coding unit (e.g., word) is interpreted within a prespecified block of material that constitutes the context

unit (e.g., sentence or paragraph). Thus, a coding variable might specify the number of times a specific word (coding unit) appears in the text, but whether the presence of a word is appropriate for inclusion in coding may necessitate that the coder reads the sentence or paragraph to understand the context. In a content analysis of song lyrics by rap artists, the researcher may be interested in the word “praying” (coding unit) and decides to design a variable that requires counting the frequency of the religious word “praying” (coding variable) in each song. However, the interpretation of the context unit of a line, a paragraph, or an entire set of lyrics may be needed to understand whether each occurrence of “praying” pertains to religion. Blindly counting the frequency of the word “praying” is misleading as some song lyrics might refer to “praying mantis” or “praying for time” or “praying to the porcelain god” – none of which would be appropriate for the religious praying variable. In many content analysis studies, context matters.

Word and theme coding variables. As content analysis of text-based information is popular in this form of research, many of the examples we discuss will involve this type of variable. A *word variable* concerns the occurrence of a particular word, synonyms of that word, or a set of words of interest. From the standpoint of identifiability, the use of a word as the coding variable represents the simplest coding strategy for textual and auditory information. An interesting example of the use of the word as a coding variable and doing a frequency count of that variable was provided by Mosteller and Wallace (1964). The investigators used content analytic techniques to determine whether the authorship of 12 *Federalist Papers* was attributed either to James Madison or Alexander Hamilton. In a preliminary investigation of Madison’s and Hamilton’s known writings, the study identified 265 words that both authors used but with varying frequencies. These known frequencies of word usage by each person were compared with those of the same words appearing in the disputed papers. By this frequency-of-use comparison between these two persons, the analysis suggested that Madison, not Hamilton, authored the papers.

A *theme variable* involves coding for underlying ideas, assertions, or perspectives. A content analysis that involves both word and theme variables may provide more information than analyses based only on words. A good example of the combination of content variables is found in the work of Stone and Hunt (1963), who attempted to determine whether they could discriminate between real and fabricated suicide notes through content analysis. They collected real suicide notes from court records and asked 15 people, matched on general demographic characteristics with the victims of these suicides, to write suicide notes that could be perceived as real. Stone and Hunt (1963) were able to discriminate between the real and the simulated notes based on three criteria. First, the real suicide notes were more likely to contain references to concrete entities (persons, places, or things); second, the real notes were more likely to use the word “love”; finally, fake notes were more likely to elaborate on thought processes and decisions about committing suicide.

The researchers arrived at these findings by counting the occurrences of these three themes and words in each of the notes, which yielded three coded variables that were subsequently used in statistical analyses. Subtracting the scores obtained on the first two variables from the third, Stone and Hunt were able to statistically discriminate between real and simulated suicide notes in 13 of 15 instances. To cross-validate their coding system, they then applied the same three criteria to another 18 notes and correctly differentiated real from simulated suicide notes in 17 of 18 instances. This “hit ratio” was significantly better than that of independent raters who were not privy to the content coding scheme. The combination of word and theme as coding variables in this study allowed for a more comprehensive extraction of the results.

The coding of themes is not restricted to textual content. The themes of underlying ideas, assertions, and perspectives can be evaluated in images and photographs. A content analysis

investigated the most prevalent themes found in memes about vaccinations by evaluating the combination of image and text for each meme (Harvey et al., 2019). Every vaccination meme was coded in terms of which of 18 possible categories of a variable best described the theme for each meme. The results found that the most common themes for the provaccine memes concerned diseases preventable by vaccines (23%), logical fallacies of antivaccine groups (13%), and vaccine safety/injury (10%). In contrast, the most prevalent themes for the antivaccine memes focused on vaccine safety/injury (28%), conspiracy theories (20%), civil liberties (12%), and informed decisions (12%).

The theme variable also can be analyzed using multimedia content. A study performed a content analysis of sexual education videos found on TikTok (Fowler et al., 2022). TikTok is a social media platform where users create and share video clips that are no longer than one minute in duration. Video clips about sex were sampled in the hashtag categories of #sexeducation and #healthclass. Afterward, the video clips were coded and classified in terms of the major sexual education theme: female and male anatomy (37%), sexual pleasure such as arousal and intimacy (23%), general sex education such as criticism and sex positivity (18%), contraception (13%), sexual health such as sexually transmitted infections (5%), and sexual communication such as parent-teen communication and providing sexual consent (4%).

Spatial and temporal coding variables. Two other types of measurement variables used in content analysis deserve mention. These include the *spatial variable*, which refers to measurement in terms of geographical, geometric, and dimensional properties (e.g., inches of newspaper column, dimensions of magazine article, shape of a photograph, number of paragraphs devoted to a topic in a newspaper article), and the *temporal variable*, which refers to measurement in terms of properties of time. Examples include minutes of a television or radio broadcast, minutes devoted to uplifting parts of presidential speeches, and seconds of sexualization in alcohol commercials found on television. Measures of this type often are used to study the degree of attention devoted to a medium for some specific type of information. For example, suppose one wished to investigate whether the President of Mexico was mentioned positively or negatively on the front pages of a sample of North American newspapers. In this project, the entire front-page articles of newspapers about the President of Mexico would constitute the *context unit*, and whether the President of Mexico is discussed in either a positive or negative light in the article represents the coding variable. The spatial variable would be precise measures of space, such as the number of sentences or paragraphs devoted to the discussion of the President. Because this is not audiovisual information stretching across time, a *temporal variable* is not pertinent to that study. However, if a content analysis involved examining video clips of anti-methamphetamine public service announcements on YouTube or TikTok, the researcher might code for the number of seconds or percentage of the video clip that was devoted to showing the undesirable consequences of using the substance. By considering all four types of coding variables – word variables, theme variables, spatial variables, and temporal variables – researchers have a greater variety of constructs for subsequent statistical analyses.

Context units. Words as the coding variable generally involve the enumeration of the occurrence of the word in the material under study. The context in which the word appears may be irrelevant to the analysis. Computers can be used to search and count the occurrence of specific words. With more complex coding units such as theme variables, some consideration of the context unit usually is required to allow confident interpretation of the meaning of the coding unit. For example, suppose an investigator was interested in studying a writer's attitudes toward communism and encountered the following sentence (see Danielson, 1963, p. 188): "The communists are taking over the world bit by bit." How is this sentence to be judged? It is impossible to code for a theme without some knowledge of the context (the paragraph or passage) in which

the sentence was embedded. If this quotation had appeared in a speech given by Vladimir Putin, it would undoubtedly be seen as a positive reference to communism.

The context unit usually is prespecified by the researcher. It defines “the largest division of context which may be consulted by a coder ...to assign a score to a basic coding unit” (Danielson, 1963, p. 188). When coding images and multimedia sources, the wider context also ought to be taken into consideration. For example, continuing with our methamphetamine memes content analysis, a researcher might be coding for behavioral action depicted in the image of each meme (e.g., sitting, standing, walking, running, or jumping). However, it is also important to evaluate that behavioral action variable in the context of the entire meme. When coding for a meme depicting the action of jumping, it would matter if the person is shown jumping off a cliff, jumping for joy, or jumping to reach a hidden stash of methamphetamines. In content with embedded meanings, the coding units and context units that are used are seldom the same. The context unit, of course, can never be smaller than the coding unit: When the theme is the coding variable, its context unit usually entails more extensive amounts of content information.

Sampling

In content analysis, the term “sampling” does not accord with the traditional definitions, which involve selecting a subset of participants from a targeted population of people. Instead, sampling involves obtaining a subset of content units from the targeted population of informational content (Neuendorf, 2011). In the techniques and approaches described in previous chapters, the number of participant units is represented as N . However, in content analysis, the number of content units is represented as k (number of sampled units of content). The statistical power to attain significance in quantitative content analysis is based on the k number of sampling units. For example, a traditional study containing 200 participants possesses the same statistical power as a content analysis study containing 200 newspaper articles, memes, websites, video clips, online message postings, and so forth.

A researcher, for example, might be interested in conducting a content analysis of racist websites hosted by hate groups. All such hate group websites would constitute the target population of relevant elements in the investigation, but because potentially thousands of these websites exist on the Internet, reviewing and content coding every one of these websites would be practically impossible. Depending on the desired degree of generalizability of findings, the researcher might select a random or non-random sampling approach by choosing 100 websites on this topic for the content analysis.

Decisions concerning the way the sample of messages is chosen for content analysis are closely related to the analyst’s choice of coding and context units. Such sampling usually involves several stages. In the first stage, the population of content units from a source is identified. Depending on the research problem, the extensiveness of this universe can vary greatly. For example, before one could study the degree of attention American newspapers devote to recurrent turmoil in the Middle East, several sampling decisions would have to be made. First, the researcher must define the population of units. Should the sample be drawn from all possible newspapers published in the United States? This limits the field somewhat because there are many good papers originating in other countries. Suppose the researcher decides to limit the sample further by examining only U.S. daily newspapers written in the English language. Should all such newspapers of this type (and there are many hundreds that meet these criteria) form the targeted population of units? Perhaps, but this would mean that those daily newspapers with circulations of 10,000 or less would be placed on par with daily newspapers with readership numbers in the hundreds of thousands, even though the influence of the former on mass public opinion is certainly less than

that of the large circulation newspapers. Because the number of newspapers with huge circulations is not great, a random sample of all U.S. newspapers would contain an overrepresentation of the smaller papers. To avoid this problem, the researcher may consider only those newspapers with circulations greater than 60,000.²

Source and content sampling. After deciding on the newspapers to be used, decisions then must be made about the extent of context to be investigated. Surely, the researcher does not wish to read the entire edition of every paper in search of Middle East news items. Coding time considerations alone would prohibit such an approach. In response to this problem, the investigator may decide to sample only front-page news, ignoring the other sections of each newspaper. This would not seem to be a particularly bad choice because the major news events of the day are almost invariably noted on the front pages of most newspapers. However, many newspapers reserve the second page for foreign news; thus, the investigator decides that the first and second pages will be searched for relevant articles about the Middle East.

Another question concerns the time period to be sampled. Is every possible edition of each sampled paper to be investigated? This will be difficult because many of the papers in the sample will probably have been established many years ago. Suppose, then, that the investigator chooses to sample only those editions published after the year 2015. Given that 365 front pages are to be reviewed every year for every selected newspaper outlet, the magnitude of the coding task may be daunting. Thus, a final sampling strategy might be applied. It involves sampling days of news copy. Rather than investigating every issue every day, each daily newspaper might be sampled every second, third, or fourth day. Or, better yet, within any given seven-day period, one, two, or three issues might be randomly selected from each newspaper. A schematic representation of the decision points that our hypothetical investigator encountered in generating this sample is presented in Figure 14.2. This figure illustrates an important distinction between two discrete types of sampling processes: source and content sampling. Decision points 1–4 are concerned primarily with the definition of the source sample from which units of content are to be drawn. Was the paper published in the United States? Was it an English language paper? Was it a daily newspaper? Was its circulation greater than 60,000? All of these questions are concerned with the source of information. The next phase is concerned with the sampling of content from within this chosen source sample. At this stage, coding and context units are sampled from each source and entered into the analysis.

The extensiveness of the content area from which messages are to be sampled, the coding and context units to be employed, and the descriptive or inferential nature of the research – all these considerations enter into the sampling decision. Not all content analyses involve a series of sampling decisions as extensive as those presented here. Consider, for example, the rather limited problem facing the investigator wishing to examine the physical qualities of the heroes of Ernest Hemingway's novels. The population of seven Hemmingway novels is not extensive. In this instance, sampling within this targeted population of sources would tend to unnecessarily restrict the raw data on which the analysis is to be based. This point concerns the issue of generalizability or external validity. Can the results be extended to Hemingway's other writings that were not analyzed (short stories, poems, newspaper articles)? In content analysis, we are concerned with representativeness. If we draw a nonrepresentative sample of the content on which our analysis is based, it is likely that our results will not be representative of the more general population of available content. A balance must be struck between efficiency and representativeness.

Often, an even more restricted selection of text, consisting of a single speech, message, or diary (Goldman & Crano, 1976), sufficiently represents the population of units. In such instances, the entire population of units (census) is selected for the content analysis. If a content analysis was performed on the literary style of books authored by Harper Lee, it would involve the only

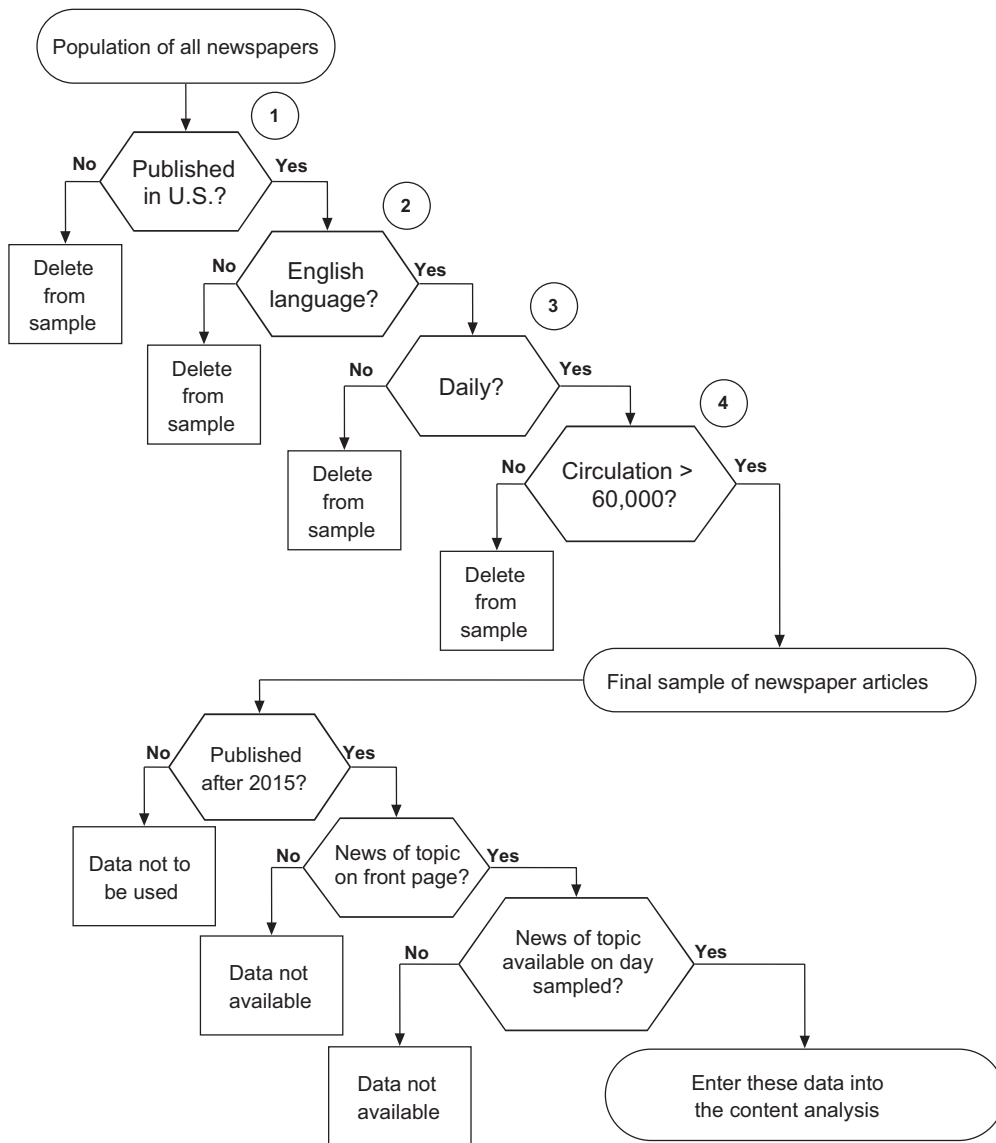


FIGURE 14.2 Schematic of decisions made in content sampling from newspapers.

two novels she ever wrote: *To Kill a Mockingbird* and *Go Set a Watchman*. The generalization based on the analysis of a single message usually is valid only when directed exclusively toward that message or communication source. To generalize to the targeted population of communications, one must either content-analyze that population through a census of all available units or randomly sample from it for analysis. Thus, it would be foolhardy to attempt to generalize the results of the content analysis to all newspapers or all novels published over the last ten years.

Use of archives. The sampling process to obtain material for a content analysis is not always as difficult as it might seem on first consideration. Usually, an investigator does not have to locate years of newspaper hardcopies and then physically wade through every issue. Thanks to computerized databases and the Internet, researchers nowadays have access to many information

sources for a content analysis. Almost all the most widely circulated newspapers can be accessed online, and search and abstracting programs can be found using simple point-and-click commands. *Lexis-Nexis* is the most widely used database of newspaper archives and is accessible through the library website of many colleges. It contains a huge archive of newspaper reports and a powerful search function. For example, it contains the full text of every *New York Times*, *Washington Post*, and *Los Angeles Times* article from 1980 onward.

The list of newspaper outlets archived in *Lexis-Nexis* could be extended almost indefinitely – the *Claremont (C.A.) College* has been abstracted by this service from September 1998, and the *Columbus Dispatch* from the first day of 1992. Broad searches could be conducted with relevant results shown regardless of the newspaper source. The search could be limited to specific newspapers and further narrowed to specific publication dates. Also available to researchers are Internet search engines (e.g., Google) to help find websites that contain archives of photographs, pictures, memes, audio clips, music lyrics, videos, television episodes, and other types of communication content that provide materials amenable to content analysis.

Coding Systems and Software

To achieve scientific respectability, the coding procedures established in the content analysis must be systematic. Documenting systematic procedures allows other researchers to replicate and extend the results of a particular content analysis study. One of the major ways of introducing systematization to this area is using prespecified classification systems in content coding. A criticism of this approach is the failure to generate mutually agreed-upon systems of coding through which diverse content could be investigated and compared across all types of content analysis projects. Researchers appear more intent on individually tailoring coding schemes to fit their research problems than on developing more generally employable techniques. Because the number of systems in the content analysis literature is so extensive, it is likely that investigators willing to search through journal articles involving content analysis in their area of research will be able to obtain a previously established coding system.

Human coders are susceptible to coding fatigue and inconsistency, but these issues do not occur when using content analysis software (Alex & Zuell, 2000; Kondrack et al., 2002). Most published studies involving content analyses do not use software programs. However, programs are available to assist with content analysis. Software to help perform a content analysis may be needed for a large-scale project that involves the evaluation of tens of thousands of units, which may take a long time for human coders to evaluate. Basic software programs tend to be more helpful for manifest variables that typically involve counting occurrences of words and simple themes and don't require the interpretation of deeper meanings by humans.

A feature of some of these basic programs is their capacity to account for synonyms when counting word occurrences, using an integrated thesaurus to perform frequency counts of an exact word only or selected synonyms of that word. Because such programs are sometimes limited in extracting subtleties in embedded meanings when interpreting themes in latent variables, problems may arise in accurately accounting for narrative structure or context. For example, basic software programs may have problems differentiating among the usage of “mean” to refer to a nasty person, the arithmetic mean, living below one's means, or one's intention to convey an idea clearly. Context matters in terms of how to properly code “mean” – a simple frequency count would not be meaningful; rather, a human coder would be necessary to decipher the word's intended meaning as operationally defined by the researcher. Typically, after the software has performed the content analysis, it is advisable to review and inspect the output to ensure that the coding rules were followed as intended by the researcher.

Perhaps the most commonly used basic commercial software for content analysis in the psychological and behavioral sciences is Linguistic Inquiry and Word Count (LIWC; <http://www.liwc.app>). This program is affordable, relatively easy to learn, and has been used in many published studies. This website contains a listing of free basic content analysis software (<http://www.tapor.ca>). Furthermore, advanced programs that use natural language processing and machine learning to help perform a content analysis are available (<https://www.g2.com/categories/text-analysis>). However, these advanced programs are expensive for most researchers and probably more affordable for companies, businesses, or entities that wish to perform many content analyses or large-scale content analyses. For example, a business may wish to search through and review the voluminous number of posted messages in online message forums, such as Reddit, regarding a company's product. Due to the implementation and use of natural language processing and machine learning, these programs are better equipped to code correctly for the content variables that take into consideration the context of the sentence of the paragraph.

Statistical Analysis

After the raw unstructured materials have been content coded into manageable quantitative coding variables, the final step to facilitate the interpretation of the rich data involves statistical analysis. Many content analysis studies simply report descriptive statistics, such as the percentage coded in each of the categories for a particular variable or the mean score for coding variables on a quantitative continuum. Other studies, aside from reporting descriptive information, provide statistical analyses to assess statistically significant associations among variables. For example, returning to our methamphetamine content analysis, to evaluate if memes depicting decaying teeth versus those that do not have that characteristic are rated higher on the quantitative variable of humorous tone. Furthermore, other statistical techniques, such as multiple regression analysis, may be conducted in which the predictors of decaying teeth, skin lesions, and premature wrinkles are used to explain the outcome of the humorous tone of the meme.

Summary of the General Paradigm

A summary of the research paradigm commonly used in content analysis studies will be useful. Not surprisingly, the series of operations to be described coincide closely with those discussed in Chapter 13, which was focused on systematic observational techniques. The scientific analysis of content usually involves the use of a prespecified coding system. The choice of the coding system is best made based on information relevant to the data to be categorized. Applied to content analysis, this suggests the researcher must become thoroughly familiar with the source or medium of the content in advance of the choice or construction of a coding scheme. Only then is the researcher able to make an informed decision about the most appropriate classificatory system. Closely bound to the choice of coding scheme are decisions about the appropriate units of analysis and the way these units will be sampled from the larger universe of potential sources. Once the coding scheme, units, and sampling rules have been decided on and coders trained, the content analysis can begin.

Coding content is much like coding observed behaviors, except the data under investigation are usually in recorded or archived form.³ As in all observational research, care must be taken to assure the reliability of the rules (or their application) used to code the variables. Krippendorff (1980) has distinguished among three approaches to reliability used in content analysis research. The simplest of these methods is concerned with *stability*, the extent to which the same coder assigns identical scores to the same content when he or she reviews it more than once in different

sessions. Failure to obtain adequate “test-retest” reliability (stability) indicates intra-observer inconsistency; the analyst, that is, did not apply the coding criteria consistently from one session to the other, suggesting a lack of adequate coder training or a coding system that is complicated or vague, and thus difficult to employ in a consistent fashion. The stability of scoring is the minimal form of reliability and should not be the sole means of demonstrating the reliability of one’s content analysis. This is because stability is dependent on the actions of a single coder, who might apply the coding rules consistently, but in an idiosyncratic fashion. Different results might be obtained if another coder attempted to use the same rules.

A more stringent approach to reliability in content analysis is *reproducibility*, which reflects the extent to which the outcome of a specific coding process is the same when different coders analyze the same content. This approach is mostly used when establishing the reliability of a coding scheme. Generally, it is termed “intercoder agreement” or “interrater reliability” because the correspondence between two (or more) coders’ estimates of the same coding variable is the measure of this form of reliability. When coding fails to generate reproducible scores for the same variable across judges, the failure can be attributed to intra-observer inconsistencies or inter-observer disagreements. Cohen’s kappa, Fleiss’ kappa, and the intraclass correlation are the recommended statistics to assess intercoder reliability (Chapter 13).

Krippendorff (1980) defines the strictest form of reliability analysis as one that compares the accuracy with which a coding system reproduces, or compares to, some known standard. For example, in Stone and Hunt’s (1963) suicide note research presented earlier, the investigators knew which of the notes were real and which were fictitious. To the extent that the coding system accurately reproduced this distinction, it was reliable (or accurate, as this form of reliability is termed). Accuracy can be weakened by intra-observer inconsistencies (poor stability), inter-observer disagreements (poor reproducibility), and the failure of the coding scheme to reflect the standard that is to be used in evaluating the content or construct of interest.

In our terms, this last form of “reliability” is a validity operation insofar as the investigation involves the correspondence between a predictor (the coded values) and an outcome (the behavior being predicted). It is a reliability operation in the sense that if intercoder reliability is deficient, the predictor will not adequately anticipate the outcome, but this same observation can be made of any measure. The unreliability of a measure (or coding scheme) inevitably diminishes validity.

Representative Examples

Following the lead of Berelson (1954) and Holsti (1968, 1969) on the variety of studies that can be conducted using content analysis, we will describe this line of research in accordance with a classic question commonly used when examining communications: “Who says what, to whom, how, and with what effect?” (Lasswell et al., 1952, p. 12). This captures the five types of topics and substantive variables (“who,” “what,” “to whom,” “how,” and “with what effect”) that may be considered when developing a content analysis study. A content analysis investigation can examine some or even all these types of variables in the same study.

Who?

“Who” refers to “who said it?” – the identity of the person or source responsible for creating the content, especially when the author or creator is unknown or disputed. Some of the most interesting studies employing content analysis have been conducted to determine the likely authorship of documents with previously unknown authors. Earlier in this chapter, we discussed

the examination of the probable authorship of 12 disputed *Federalist Papers* (Mosteller & Wallace, 1964). In an investigation similar in form, Yule (1944) was able to point quite convincingly to the probable author of the classic *The Imitation of Christ*.

These studies do not always provide unambiguous results. Biblical scholars, too, have used content analytic techniques in attempting to settle questions of disputed authorship. For example, Morton (1963) identified seven elements of writing style that he felt would clearly distinguish among authors.⁴ Using this system, he concluded that six different authors were responsible for the epistles traditionally ascribed to St. Paul and challenged the orthodoxy of Christianity either to debunk his results or to revise their traditional views of the early church. The challenge was met by Ellison (1965), who, employing Morton's seven indicators, "discovered" that James Joyce's *Ulysses* was written by five different authors, as discerned by passages that appear to be stylistically different in a content analysis, and concluded that none of these individuals could have written *Portrait of the Artist as a Young Man*! The obvious failure of Morton's previous coding system to later identify the single authorship of *Ulysses* and *Portrait of the Artist* raised doubts about the validity of that coding system.

What?

"What" refers to examining and understanding "what was said?" Topics and variables of this type focus on the characteristics of the substantive message used in the content. Many content analyses have focused on consideration of the research question of what is the message or information embedded in the content. Berelson (1954, pp. 490–495) provided a classificatory system to study variables in content analysis focused on the "what."

Trends. The study of communication trends over time is a research possibility. Has the level of achievement motivation expressed in the stories included in children's storybooks changed during the last 50 years? Is the release of political movies associated with certain economic periods? Are the values and goals expressed in the movies today the same as those of 20 years ago? Is self-disclosure of information higher today on social network websites than compared to a decade ago? Content analyses that examine trends assume that the mass media serves as a barometer of various aspects of societal trends, customs, and values.

Yakobson and Lasswell (1949), for example, studied the content of Russian May Day slogans from 1918 to 1943. During that period, political scientists noted a gradual mellowing of revolutionary zeal on the part of the Soviets. Consistent with this view, a content analysis of slogans revealed that calls for universal revolutionary activities had steadily diminished over time and were replaced with an increased emphasis on nationalistic appeals.

In another example of the examination of "what?" that focused on societal trends, Christodoulou and Lac (2021) reviewed electronic cards (e-cards) that could be sent to congratulate the parents of newborn infants. The purpose of the content analysis was to investigate how sex roles and gender stereotypes were communicated and perpetuated through electronic birth congratulatory cards in modern society. The sample of 200 congratulatory birth cards was obtained from the ten most popular e-card websites that were listed first after an Internet search. Each electronic card was classified as for the parents of newborn girls or boys. The findings of this electronic card study (Christodoulou & Lac, 2021) were compared to a physical card study about birth congratulation cards published about two decades earlier (Bridges, 1993). In both the electronic card and physical card studies, the cards for newborn girls contained more stereotypically feminine images of frills/lace/ribbons and floral decorations, and the cards for newborn boys contained more stereotypically masculine images of geometric and plain/striped decorations. However, the physical card investigation found greater mentions of words and synonyms of

“sweet,” “sharing experiences,” and “happiness” in the girl compared to boy cards, but the electronic card investigation did not reveal such gender differences. These two content analyses suggest that some gender stereotypical messages have persisted during the past two decades.

Norms. Research applying content analysis could be used to understand societal and group norms. Hum and colleagues (2011) content-analyzed the main profile photo that people used for their Facebook accounts. The objective of that study was to understand the norms of how people represented their social identities online through a mere photograph. The sampling units were drawn from profiles belonging to a sample of people who were Facebook friends of the researchers. This convenience sampling technique was pursued because it would not be possible to view Facebook profiles unless those members friended the researchers or made their profiles publicly accessible to others. Data from the coded variables revealed that most profile photos depicted each person individually without others present, showed the person in a still pose and therefore not engaged in physical activity, and were found to be professionally appropriate. This study provided social normative standards regarding how profile photographs are displayed by users of this social media platform.

International differences. Another application is the study of differences in communication content occurring across nations and cultures. Similarities or dissimilarities among the content produced and disseminated by different nations or cultures are presumed to reflect important aspects of the respective countries. Consider, for example, an investigation undertaken by Lewin (1947), in which he compared the literature disseminated by Hitler Youth with that of the Boy Scouts of America. The major themes of both content samples stressed the value of physical fitness, discipline, and achievement. However, in the Hitler Youth literature, more emphasis was placed on national loyalty and racial pride, whereas the Boy Scout sample stressed generosity and imagination. This study provided support for a hypothesis regarding the child training practices that gave rise to the authoritarian Nazi regime during World War II.

Standards. Content analysis can be used to assess whether media outlets meet industry standards of unbiased coverage. Do news magazines fulfill their objectives of presenting a fair, unbiased discussion of public events? In an examination of the coverage afforded the major candidates in the 2008 and 2012 U.S. presidential elections, the Pew Research Center's Project for Excellence in Journalism found that the tone of the election coverage slightly favored the Democratic candidate over his opponents. In earlier research on the treatment of Republican candidates for the presidency, Westley et al. (1963) found that the reporting in *Time*, *Newsweek*, and *U.S. News and World Report* on the treatment of Republican leaders was only very slightly more favorable than that afforded the Democrats.

To Whom?

“To whom” refers to the intended or targeted audience of the communication. Content analysis can be used to provide empirical estimates of the extent to which communicators vary their messages as a function of the audience. For example, John Foster Dulles, Secretary of State during the Eisenhower presidential administration, was the focus of two such investigations. Cohen (1957) and Holsti (1962) found clear evidence that the content of Dulles' communications was guided by a consideration of the audience. Similarly, Berkman (1963) found that the type of product advertisements found in “Life” and “Ebony” magazines could be differentiated in terms of the socioeconomic status of the readers.

Another example of a content analysis that examined the characteristics of messages as delivered to a specific targeted audience concerned memes about vaccinations (Harvey et al., 2019). The study identified a comprehensive listing of 128 Facebook fan pages devoted to the topic of

vaccination. Out of these, the top five provaccine and top five antivaccine Facebook pages with the greatest number of followers were selected based on the justification that these fan pages would have the widest reach in the number of viewers. In each of these ten vaccine fan pages, up to 30 unique memes were sampled. After identifying the units that satisfied the criteria for a meme (text overload on an image) and were not duplicates, this yielded a total of 234 unique memes for the content analysis. Two raters coded for every meme to assess interrater reliabilities. Statistical tests showed that the antivaccine memes (intended for an antivaccine audience) were rated to be significantly higher on emotional appeals and fear appeals, but the provaccine memes (intended for a provaccine audience) were higher on sarcasm (usually directed at antivaccine views). Furthermore, the vaccine text in every meme was fact-checked against published informational papers found through reputable organizations such as the Centers for Disease Control and the National Institutes of Health. The antivaccine memes contained a significantly higher percentage of false claims (30%) compared to provaccine memes (4%). This study suggests that provaccine and antivaccine advocates differ in how their message characteristics and persuasion tactics are tailored to their in-group audience members.

How?

Investigations of “how” focus on the style form of communication. Although the same information might be presented in two communications, one message may be considerably more influential than the other because of the way the information was presented. Analysis of how messages are structured and disseminated constitutes a major goal of content analysis investigations of the “how” variety.

Propaganda analysis. Propaganda analysts have been particularly active in this area of research. Lasswell (1927), for example, attempted to identify the reasons that underscored the British propagandists’ success in World War I, and the concomitant German failure, by isolating four major goals that both sides attempted to realize. The first involved attempts to generate and maintain home front hostility toward the enemy, the second stressed preservation of friendly alliances, the third concerned attempts to secure the support of neutral countries, and the fourth focused on efforts to demoralize enemy soldiers and civilians. Because the propagandistic goals of the British and Germans were essentially the same, to gain an appreciation of the effectiveness of one side over the other, their communications had to be analyzed not in terms of *what* was said but in terms of *how* it was said. A major reason for the British propagandists’ success can be traced to the image of a beleaguered island nation they attempted to convey. The British pictured themselves as peace-loving islanders, forced to fight so that Western Civilization might be saved from the barbarians. Their goal in the war was “to end all wars,” and their methods were strictly humanitarian. German propagandists cited the need to extend Germanic *Kultur* as a justification for their war efforts. Humanitarianism was given short shrift, and British atrocities were rarely noted. Like their counterparts in the field, the German public relations staff was soundly defeated.

Propaganda also may be disseminated on the Internet through websites and online fringe groups seeking to promote risky lifestyles. Borzekowski et al. (2010) evaluated how websites promoted eating disorders. After conducting searches using 15 pertinent search terms, the content analysis included a relatively exhaustive listing of 180 pro-eating disorder websites. The content analysis showed that the vast majority (91%) of these sites were accessible to the public and did not require any special membership or password. Most of the websites presented information that was pro-anorexia (84%) and pro-bulimia (64%). Furthermore, these websites were coded in terms of the types of eating disorder tips provided using a variety of coding variables,

including dieting (74%), foods considered to be safe (68%), exercising (59%), purging (50%), pills and laxatives (49%), how to distract oneself from eating (48%), how to hide the eating disorder from friends and family members (23%), and alternative medicines (22%). Also, statistical analyses were conducted with each of these manifest variables and the latent variable of harm of the website to the public (1 = not too harmful to 5 = very dangerous). The results disclosed that the presentation of each of these tips was significantly associated with greater rated harm of the website.

Stylistic analysis. The analysis of propaganda does not exhaust the question of “how?” Miles (1951), for example, attempted to describe in a quantitative fashion the differing stylistic patterns characteristic of distinct literary periods. Her analyses of poetic writings proved quite successful. Harvey (1953) attempted to discover the major distinguishing characteristics between best-selling and poor-selling novels. In combining several content variables in a prediction equation, the study found that the variables could predict sales with better than 80% accuracy.

Furthermore, studies focused on stylistic analysis may examine the extent that a communication or message is easy or hard to understand by readers. *Readability* refers to the ease that readers can understand written text – a high readability index signifies easier and faster speed of reading the written materials for people in general. Readability formulas, such as the Flesch-Kincaid readability test, are determined usually based on sentence length and the average number of syllables per word and other characteristics such as the difficulty of the vocabulary words. Some readability indices generate information about the minimum level of education required to understand the materials, for example, a text may require a seventh grade reading level to comprehend. This website provides the calculation of various readability indices, including the Flesch-Kincaid readability test, after pasting in text (<https://www.analyzemymywriting.com>). A written passage, essay, newspaper article, textual document, or book that contains short sentences involving few words and words with few syllables tend to be judged to be easier to read by readability indices. For example, books designed for children compared to adults should yield a higher readability score. A romance novel aimed at a general audience versus a technical textbook should produce a higher readability score.

Some content analysis studies applying readability indices have evaluated the readability of questionnaire items of validated measures in a domain of research (e.g., measures about gambling). Readability indices may not always be helpful, as they are sensitive to idiosyncratic and nonstandard use of language of some writers. Gertrude Stein’s writing is characterized by short sentences and short words, but her writing is difficult to comprehend. Her poems and novels would confound almost all readability formulas that would erroneously indicate that Stein’s writings are easy to read when they are not.

With What Effect?

The final “with what effect” concerns investigations focused on the effects and consequences of a communication message on its receivers and society. For example, sophisticated correlational analyses have been conducted to assess the association between the coverage of championship heavyweight boxing matches in the media and subsequent increases in homicides in the United States (Phillips, 1983). Phillips (1982) also presented evidence based on archival records suggesting a link between the occurrence of suicide content depicted in popular television soap operas and the subsequent increase in the incidence of actual suicides in the general population. Vitaglione (2012) suggested that communication effects on people could generalize beyond suicides in research that indicated that shortly after highly publicized NASCAR races, aggressive driving accidents and injuries increased considerably.

Conclusion

The value of content analysis lies in its versatility. Information that would be dangerous or impossible to research can be studied with this approach using publicly available content materials. The technique has been used to study almost all features of the communication process, from the disputed authorship of books and papers to understanding communication as expressed on social media and the Internet. When designing a content analysis study, decisions ought to be made regarding whether it will be a qualitative or quantitative content analysis study. Another consideration is whether the study should involve a deductive content analysis or inductive content analysis. Furthermore, each variable of interest should be pondered in terms of its measurement as a manifest or latent variable. Even after resolving that measurement issue, the researcher still needs to consider whether the content unit of each variable should be assessed in the form of a word variable, theme variable, spatial variable, or temporal variable. Content analysis research is considered nonexperimental, so causal claims are not warranted usually. Content analytic methods tend to be nonexperimental and therefore not ideal for establishing causal relations, but they may be useful in developing ideas and hypotheses that can be tested in controlled laboratory environments.

Notes

- 1 This chapter focuses on content analysis for research purposes. In private companies, a content analysis and information derived from that method may help with monitoring, filtering, and screening media communications.
- 2 More than 100 American daily newspapers satisfy this criterion.
- 3 Some analyses are performed on verbal material, but the audio clips are usually printed before analysis is begun.
- 4 The stylistic indicators employed by Morton were sentence length, frequency of use of the words *and*, *but*, and *in*, frequency of use of definite articles, of all forms of the verb to be, and of use of the third person pronoun.

References

- Alexa, M., & Zuell, C. (2000). Text analysis software: Commonalities, differences, and limitations: The results of a review. *Quality and Quantity*, 34, 299–321.
- Berelson, B. (1954). Content analysis. In G. Lindzey (Ed.), *Handbook of social psychology*. Reading, MA: Addison-Wesley.
- Berkman, D. (1963). Advertising in Ebony and Life: Negro aspirations vs. reality. *Journalism Quarterly*, 40, 53–64.
- Borzekowski, D. L., Schenk, S., Wilson, J. L., & Peebles, R. (2010). E-Ana and E-Mia: A content analysis of pro-eating disorder websites. *American Journal of Public Health*, 100, 1526–1544.
- Bridges (1993). Pink or blue gender-stereotypic perceptions of infants as conveyed by birth congratulation cards. *Psychology of Women Quarterly*, 17, 193–205.
- Brubaker, P., Boyle, K., & Stephan, D. (2017). The shared cultural experience: A comparison of religious memes created by the Church of Jesus Christ of Latter-day Saints, religious media, and church members. *Journal of Media and Religion*, 16, 67–79.
- Christodoulou, J., & Lac, A. (2022). Examining the communication of gender roles to parents: A quantitative content analysis of online birth congratulation cards. *Psychology & Sexuality*, 13, 640–651.
- Cohen, B. D. (1957). *The political process and foreign policy: The making of the Japanese peace settlement*. Princeton, NJ: Princeton University Press.
- Dale, E., & Schall, J. S. (1948). A formula for predicting readability. *Educational Research Bulletin*, 27, 37–54.
- Danielson, W. A. (1963). Content analysis in communication research. In R. O. Nafziger & D. M. White (Eds.), *Introduction to mass communications research* (pp. 180–206). Baton Rouge: Louisiana State University Press.

- Ellison, J. W. (1965). Computers and the testaments. In *Proceedings: Conference on computers for the humanities* (pp. 64–74). New Haven, CT: Yale University Press.
- Fowler, L. R., Schoen, L., Smith, H. S., & Morain, S. R., (2022). Sex education on TikTok: A content analysis of themes. *Health Promotion Practice*, 23, 739–742.
- Goldman, R. M., & Crano, W. D. (1976). Black boy and child in the promised land: Content analysis in the study of value change over time. *Journal of Black Studies*, 7, 169–180.
- Harvey, A. M., Thompson, S., Lac, A., & Coolidge, F. L. (2019). Fear and derision: A quantitative content analysis of provaccine and antivaccine internet memes. *Health Education & Behavior*, 46, 1012–1023.
- Harvey, J. (1953). The content characteristics of best-selling novels. *Public Opinion Quarterly*, 17, 91–114.
- Holsti, O. R. (1962). The belief system and national images: John Foster Dulles and the Soviet Union. Unpublished doctoral dissertation, Stanford University.
- Holsti, O. R. (1968). Content analysis. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (pp. 596–692). Reading, MA: Addison-Wesley.
- Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Reading, MA: Addison-Wesley.
- Hooker, E. (1957). The adjustment of the male overt homosexual. *Journal of Projective Techniques*, 21, 18–31.
- Hum, N. J., Chamberlin, P. E., Hambright, B. L., Portwood, A. C., Schat, A. C., & Bevan, J. L. (2011). A picture is worth a thousand words: A content analysis of Facebook profile photographs. *Computers in Human Behavior*, 27, 1828–1833.
- Kondracki, N. L., Wellman, N. S., & Amundson (2002). Content analysis: review of methods and their applications in nutrition education. *Journal of Nutrition Education and Behavior*, 34, 224–230.
- Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Beverly Hills, CA: Sage.
- Lac, A. (2016). Content analysis. In R. J. R. Levesque (Ed.), *Encyclopedia of adolescence* (2nd ed., pp. 1–5). Switzerland: Springer International.
- Lasswell, H. D. (1927). *Propaganda technique in the world war*. New York: Knopf.
- Lasswell, H. D., Lerner, D., & Pool, I. de S. (1952). *The comparative study of symbols*. Stanford, CA: Stanford University Press.
- Lewin, H. S. (1947). Hitler youth and the Boy Scouts of America: A comparison of aims. *Human Relations*, 1, 206–227.
- Martínez-Rolán, X., & Piñeiro-Otero, T. (2016). The use of memes in the discourse of political parties on Twitter: Analysing the 2015 state of the nation debate. *Communication and Society*, 29, 145–159.
- Miles, J. (1951). *The continuity of English poetic language*. Univ. of California Publications in English. Berkeley: University of California Press.
- Morton, A. Q. (1963). A computer challenges the church. *The Observer*, November 3, 21.
- Mosteller, F., & Wallace, D. L. (1964). *Inference and disputed authorship: The Federalist*. Reading, MA: Addison-Wesley.
- Neuendorf, K. A. (2011). Content analysis – A methodological primer for gender research. *Sex Roles*, 64, 276–289.
- Phillips, D. P. (1982). The impact of fictional television stories on U.S. adult fatalities: New evidence of the effect of the mass media on violence. *American Journal of Sociology*, 87, 1340–59.
- Phillips, D. P. (1983). The impact of mass media violence on U.S. homicides. *American Sociological Review*, 48, 560–568.
- Shifman, L. (2013). *Memes in digital culture*. Cambridge: MIT Press.
- Stemler, S. (2001). An overview of content analysis. *Practical Assessment, Research & Evaluation*, 7, 1–6.
- Stone, P. J., & Hunt, E. B. (1963). A computer approach to content analysis using the general inquirer system. In E. C. Johnson (Ed.), *American Federation of Information Processing Societies, conference proceedings* (pp. 241–256). Baltimore, MD: Association for Computing Machinery.
- Vitaglione, G. D. (2012). Driving under the influence (of mass media): A four-year examination of NASCAR and West Virginia aggressive-driving accidents and injuries. *Journal of Applied Social Psychology*, 42, 488–505.
- Yule, G. U. (1944). *The statistical study of literary vocabulary*. London: Cambridge University Press.
- Westley, B. H., Higbie, C. E., Burke, T., Lippert, D. J., Maurer, L., & Stone, V. A. (1963). The news magazines and the 1960 conventions. *Journalism Quarterly*, 40, 525–531.
- Yakobson, S., & Lasswell, H. D. (1949). Trend: May Day slogans in Soviet Russia. In H. D. Lasswell, N. Leites, R. Fadner, J. M. Goldsen, A. Gray, I. L. Janis, A. Kaplan, D. Kaplan, A. Mintz, I. de S. Pool, & S. Yakobson (Eds.), *The language of politics: Studies in quantitative semantics* (pp. 233–297). New York: George Stewart.

15

INTERVIEWING

A research **interview** is a data collection method in which participants verbally communicate information about their behavior, thoughts, or feelings in response to questions verbally posed by an interviewer. Unlike most of the methods discussed in the two preceding chapters, interviews always involve some form of verbal interaction in which the investigator requests information from a respondent. This feature distinguishes the technique from self-administered questionnaire methods, where respondents may never interact personally with a researcher. The interactive process of the interview, and its dependence on verbal or linguistic responses, constitutes both its major strength and its major drawback as a method of scientific social research.

It is almost always easier and less expensive to use written questionnaires than to conduct interviews (Bartholomew et al., 2000). Thus, it is important to consider circumstances when the interview method is optimal. The most important factor in choosing the interview method involves research issues requiring personal, interactive, and verbal methods of data collection. This might be the case, for instance, when highly sensitive information is sought, or when certain responses call for verbal probing for details that would be difficult to elicit using a self-report questionnaire format.

Interviews also might be required with special respondent populations who have challenges in comprehending a questionnaire (e.g., young children, the elderly, and the illiterate). Further, if the problem of nonresponse poses a serious threat to external validity, it may be more likely that efforts to cultivate rapport with the interviewee via personal contact may achieve higher response rates than less personal self-reported questionnaire approach.¹

Modes of Administration

Before beginning the discussion on designing and conducting interviews, it is important to note that interviews do not always entail face-to-face encounters between interviewer and respondent. *Telephone interviews* are common and popular, with good reason, and online interviews are now widespread in business, industry, and academic settings. Not many years ago, researchers were warned to avoid non-face-to-face interviews because it was assumed that people would be less willing than in face-to-face encounters to be interviewed, and those who did agree to

participate would be unwilling to give the interviewer more than five to ten minutes of their time. Seminal research focused specifically on this concern indicated it was much less problematic than originally thought. Groves and Kahn (1979), for example, found participation rates in telephone interviews were only 5% less than those involving face-to-face encounters, and Bradburn and Sudman (1979) demonstrated somewhat higher response rates to telephone, as opposed to personal interviews, at least in urban areas, and Dillman's (1978) study revealed the response rate in telephone interviews was approximately 15% greater than that of mail surveys. Online interviews have become increasingly popular. In hiring interviews, for example, research indicates that nearly 63% of HR managers used video interviews (see <https://vervoe.com/online-interview-guide>).

In terms of respondents' willingness to participate in extensive telephone interviews, implications are much the same as those drawn from the response rate research.² Research by Quinn et al. (1980), Dillman (1978), and Smith and Kluegel (1984) all indicates that once committed to the telephone interview, respondents are unlikely to disengage prematurely. Studies using telephone interviews lasting up to an hour have been conducted with no appreciable dropout problem (Kluegel & Smith, 1982). Herman (1977) coupled face-to-face with telephone interviewing. An interviewer went to the homes of those who refused a telephone interview or could not be reached via phone. She found that the quality of responses collected from phone interviewees was comparable to that obtained from face-to-face respondents, although the phone interviewees were less likely to disclose personal information (e.g., for whom they voted). These findings supporting the comparability of telephone and face-to-face approaches still hold today (Hajebi et al., 2012; Lee et al., 2008).

Online interviews share many of the same features as in-person interviews. The time needed for interviews depends on the nature of the issue, but somewhat more time may be needed to ensure that the technology is cooperating. Once this is assured, the interview should proceed much as in face-to-face approaches. The online method often allows for easier presentation of prompts and written questions, and probably will become the most preferred format, owing to its low expense and capability of presenting questions and illustrations efficiently and conveniently.

Despite the many advantages of the telephone approach, it is important to keep in mind some potential disadvantages. In face-to-face and online interviews, the interviewer is more likely to detect and correct confusion on the part of the respondent (Bartholomew et al., 2000). The face-to-face and online interviewer is more likely to be able to clarify issues, and to realize that a question posed does not carry the intended implication. This is an especially important advantage in the initial research phases in which the interview schedule (the list of questions to be asked) is being developed. The telephone interview also does not allow for the use of visual aids; this can prove to be important if complex questions or lengthy response options are to be provided to the interviewee. Finally, the telephone interview does not provide the researcher with visual contact with the respondent. This can prove troublesome in situations in which visual cues are used to replace otherwise lengthy or overly personal questions. In health surveys, for instance, a cursory visual inspection of the physical status of the respondent often proves useful in identifying whether the person is malnourished. Similarly, the telephone approach would seem less than optimal in situations in which respondents' socioeconomic status is to be estimated by visual inspection of their neighborhood or dwelling. In these cases, the telephone interview is not necessarily less costly than the more personal face-to-face approach. No matter what form the mode of interviewing assumes (telephone, online, or face-to-face), the principles that govern good interview technique are the same. In the pages that follow, we will present guidelines that should be followed when using this method.

Developing the Interview

The interview method can be used to study a range of issues, across widely diverse respondent samples. The methodological variations available in the choice of specific interview method, question format, and so on are equally diverse. Nevertheless, there are many common features that all interview research shares, which reflect the decisions the investigator faces at one time or another in the design of any interview study (Figure 15.1). We focus on these “choice points” in the following discussion to provide a general framework within which all the many and varied forms of interviews might be conceptualized.

The strategic research procedures characteristic of all interview studies involve decisions concerning (1) the question content, (2) the interview format, (3) how the interview will be conducted, (4) the sample or population from which respondents will be chosen, and (5) the methods to be used in coding, aggregating, and analyzing respondents’ answers. The central task of the present chapter reduces to a consideration of the first three decisions.

Question Content

Although social scientists have become increasingly aware of the necessity for a systematic study of best practices for developing interview questions, there is surprisingly little in the way of research directed specifically toward this issue. Some (Dijkstra & van der Zouwen, 1982; Krosnick & Fabrigar, 1997, 2001; Rossi et al., 1983; Sudman et al., 1996) do discuss some of the intricacies involved, but most of the advice seems to boil down to “use common sense” and “draw upon practical experience.” This is undoubtedly sound advice, but it provides little direction for researchers in terms of specific item wording or format. Later in the chapter we discuss some of the factors that appear to affect the likelihood that people will interpret the meaning of a question properly, but for the moment it is sufficient to keep in mind a few simple but effective rules of thumb in question construction:

- Keep the items as brief as possible – the longer the item, the more likely it is to succumb to one or another of the problems listed below.
- Avoid subtle shadings – if you want to know about something, ask about it as directly as possible.

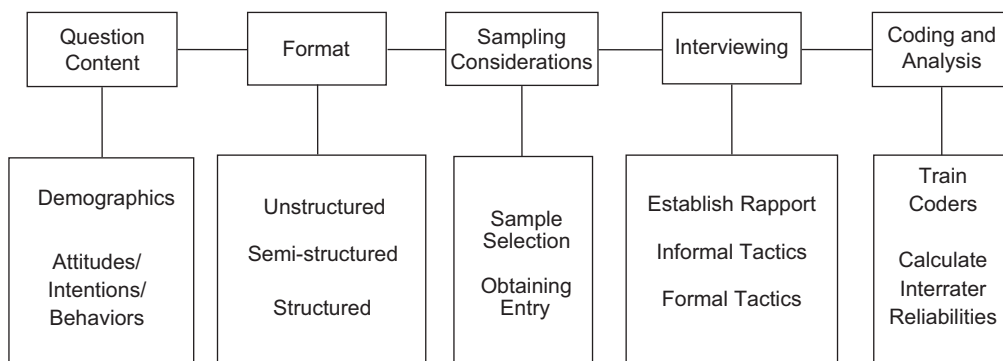


FIGURE 15.1 Steps in designing an interview study. The top row are the broad dimensions to be assessed. The connected items on the bottom row illustrate some of the features linked to the broader dimensions.

- Avoid double-barreled questions, i.e., questions that logically allow for two (possibly opposed) answers – “Do you like this year’s Fords or Chryslers?”
- Use language the respondents understand (most people are not social scientists, so the jargon of the field is ill-advised).
- If possible, pilot test the items on a small subsample of respondents drawn from the same population as the final sample.

The variety of issues that can be addressed in the interview represents one of the most appealing features of the methodology. However, this diversity makes it difficult to categorize the types of questions that have been and can be used. Some researchers use a simple dichotomy to classify survey items – does the question focus on a public or a private action? Or is the item concerned with *fact* or *opinion*? Experience suggests that these classifications are unnecessarily broad. Schuman and Kalton (1985) suggested a more differentiated classification, which more sensitively describes the situations in which the interview might be used most profitably and the forms of information that might be sought through its use. We adopt their scheme with minor modifications.

Demographic information. Questions concerned with descriptive personal characteristics of the respondent – age, religion, sex, race, gender identity, income, etc. – are perhaps the most common of all items used in interviews. In general, there is reason to believe that answers to questions of this type can be trusted, especially if the item is clear and precisely worded (e.g., Parry & Crossley, 1950; Weaver & Swanson, 1974). So, for example, it would be better to ask, “date of birth” than “age,” to determine the age of a respondent. The former is readily available in most people’s memories and is less likely to be distorted because of social desirability bias.

Two demographic variables that sometimes defy this observation (for rather different reasons) are race (ethnicity) and income. Given the subjective nature of race (even social demographers are hard put to provide an acceptable categorization system), and the even more subjective nature of ethnic origin in the United States, where the “melting pot” analogy is far more than symbolic, it is difficult to validate participants’ responses to items of this type. Many view income as a private matter, and thus, questions on this issue typically generate a relatively large (on the order of 51%) refusal rate. Nonresponse (or item refusal) is preferable to deliberate misreporting by a respondent, but it does cause problems in later analytic stages of the research, as refusal is usually treated as a missing response in an analysis, which can cause some statistical programs to delete the entire case by default (see Chapter 12 for information on handling missing data). In addition, validity is cause for concern even with respondents willing to disclose personal income. For many, income varies as a function of the availability of overtime, seasonal fluctuations in worker demand, etc. In other circumstances, especially when total household income is at issue, systematic underreporting is likely because of income sources that typically are overlooked (e.g., interest, dividends, irregular or seasonal employment), producing a less than complete knowledge of household members’ wages. Thus, estimates of income are likely to be imprecise.

With such problems, it is little wonder that some interview questions now seek the less threatening information regarding the respondent’s job, and then attempt to relate these answers to other items, in extrapolating income or socioeconomic status (SES) from them (Cain & Treiman, 1981; Hodge et al., 1964; Hollingshead, 1949). Given the near ubiquity of such items in almost all surveys, it is understandable why researchers have recommended that a standardized set of demographic items be used in all interview research. We oppose such a recommendation, because it represents a move in a direction away from the tailoring of questions to research objectives, a central tenet of good research design. For example, in examining socioeconomic

status, the objective of one study might be to measure the concept using annual salary, whereas another study might seek a measure of total assets.³

Reconstruction from memory. The interview is a practical means of investigating people's recall of past events. Some events having important social implications occur so rapidly or unexpectedly that researchers are unable to observe behavior at the time the events occur. Participants' recall of floods, blizzards, or riots, for example, might provide valuable information for the social scientist. The investigation of less encompassing, but perhaps equally important past events at a personal level (e.g., marriage, births, promotions, etc.) also can be a rich source of information, but often they leave few accessible traces. By judicious use of the interview, such information becomes more available for research.

This is not to say that the data about past events obtained through interview techniques are necessarily trustworthy. Some events simply are not important enough to the respondent to remember. Consistent with this observation, Sudman (1980) has speculated that there are three important factors that influence fidelity of recall. They are:

- *Uniqueness of the event.* Did it stand out from other events? (e.g., most older Americans can remember exactly where they were, and what they were doing, when they learned that President John Kennedy had been shot).
- *Magnitude of the event.* Did it result in major economic or social costs or benefits? We're more likely to recall the day we won the \$1 million lottery than the time we won \$12 at the racetrack.
- *Long-term impact of the event.* Did it produce long-lasting consequences? Recall of an injury that led to amputation will be more memorable than one that, while equally serious, had consequences of limited duration.⁴

Although the specific event under study will have a major influence on the likelihood of recall accuracy, the nature and form of the questions used to induce participants to reconstruct memories in interviews also can have a significant effect on accuracy. Cannell et al. (1981) suggest instructions that stress the importance of the interview topic positively affect the completeness of people's responses. Indeed, Cannell and associates sometimes go so far as to ask the respondent to sign an agreement acknowledging the importance of accuracy and completeness of answers.

Also effective in such situations is the use of longer, rather than shorter, questions. People tend to give longer and more complete answers to longer questions, even if the longer question is merely a multiple restatement of the shorter version (Ray & Webb, 1966). This is likely attributed to the implicit social norm of reciprocity: When you ask someone a wordy question, the other person will likely offer a wordier response. Although briefer questions are generally preferred, longer questions may be used when accuracy of recall is at stake. Finally, by reinforcing more complete answers, the interviewer can encourage respondents to provide more extensive, and presumably more accurate, responses. Although there is some danger that such reinforcement could bias participants' responses, if used judiciously this tactic appears capable of stimulating more complete answers without unduly influencing their content.

Attitudes, intentions, and behaviors. Interviews are commonly used to assess people's attitudes, intentions, and behaviors. We distinguish between these terms because an attitude is an evaluative belief about a person or thing (e.g., "How positively or negatively do you evaluate the present Supreme Court?"), which may or may not carry with it any behavioral implications, whereas an intention is an indication of a person's decision to act (e.g., "How much do you like to exercise?" versus "How many times do you plan on exercising in the coming month?"). The items tap types of beliefs that are conceptually different from behavior ("How many times did

you exercise in the past month?”). In Chapter 16, we discuss the elements that must be considered in developing attitude measures, and this information will not be repeated here.⁵ However, Schuman and Kalton (1985) have identified two aspects of interview questions – constraint and specificity – that are particularly relevant to interviews in which people’s attitudes, intentions, and behaviors are the focus of study.

Question specificity refers to how the question is phrased, implicating that minor changes of wording may have marked effects on people’s responses. Rugg (1941), for example, found that only 25% of respondents in a national sample were willing to “allow speeches against democracy,” but, when the question was put to a comparable sample in slightly modified form, Rugg found that 46% were against “forbidding” such speeches. Evidently, the different connotations of the words “forbid” and “(not) allow” produced these differences, even though the implications of the two questions were identical.

Question constraint refers to the response format and the options available for answering a question. *Open-ended questions* allow respondents to supply any relevant response (“What is your favorite way to exercise?”); *closed-ended questions* restrict the options that can be chosen in answering the question (Which is your favorite way to exercise? Walking, Running, Biking, Lifting Weights...?). Respondents’ answers are constrained by the available choices.

A common constraint involves participants’ choice of possible responses that do not include the “Don’t Know” or “No Opinion” option. Of course, if a respondent spontaneously chooses one or the other of these “don’t know” responses, it typically is recorded as such. It would be foolish to force an answer when the respondent indicates neither knowledge nor interest in the topic. However, Schuman and Presser (1981) demonstrated that the mere presence of an explicit “no opinion” option has a strong effect on the proportion of respondents likely to answer in this manner. Questioning respondents under standard conditions about an obscure piece of legislation (the Agricultural Trade Act of 1978), the researchers found that two-thirds of their sample had “no opinion.” When an explicit “no opinion” option was made available to another, comparable respondent group, the proportion that chose this alternative rose to 90%.

Given this apparent effect of the “don’t know” option, researchers should evaluate whether its use will increase or decrease the accuracy of the information obtained. If the issue involves factual information and the researcher judges that the respondent population includes many people having no knowledge of the issue at hand, then it is advisable to include this option (Bishop et al., 1980).⁶ Conversely, if the measure is concerned with beliefs, attitudes, or values, and the researcher judges that most people do have opinions on the topic (though it might not be altogether firm), the “don’t know” option should be avoided as it can result in considerable underreporting of opinions (Krosnick et al., 2002). A good practice is to avoid disclosing to participants that there is a “don’t know” option – and allow them to spontaneously offer this answer, if indeed they do not possess an opinion about the matter. The “don’t know” option is undesirable because such answers usually are treated as missing values in most statistical analyses. This compromises the statistical power in many statistical techniques, especially ones using listwise deletion, in which *all* responses for the participant choosing the *don’t know* option are discarded when even one “don’t know” answer is given.

Sensitive beliefs and behaviors. Sometimes, an investigation may be focused on highly personal, secretive, or illegal behavior. People engaging in such actions usually are unwilling to be observed when doing so; however, surprisingly, they often are quite willing to discuss their experiences, especially if they can be assured of confidentiality. Studies of sexual practices (Kinsey et al., 1948), marijuana usage (Schaps & Sanders, 1970), drunken driving (Bergen et al., 2012) are examples of interviews focused on actions people would not like to be observed doing.

Considerable research has investigated the degree to which respondents underreport the extent of their involvement in private, embarrassing, or socially undesirable actions. The research on this issue is mixed. In some areas, research suggests the validity of self-reports of illicit activity. For example, considerable research suggests that with appropriate safeguards, adolescents' self-reports of drug use are trustworthy (Denis et al., 2012; Hornik & Orwin, 2012; Richter & Johnson, 2001; Smith et al., 1995). Testing approaches that include a biomarker along with a self-report generally report somewhat lower levels of response validity and sensitivity (Magura, 2010). In studies with less strenuous methods to guarantee truthful reporting, findings indicate a tendency for people to distort or fail to admit to actions that are illegal or viewed with disapproval by the society at large. For example, married respondents underreport sexual infidelity (Andrews et al., 2008), their BMIs, especially if they are obese (Visscher et al., 2006), criminal recidivism (Kroner et al., 2007), use of cocaine and marijuana during the third trimester of pregnancy (Bessa et al., 2010), and past histories of sexually transmitted diseases (Hong et al., 2011).

Attempts to overcome underreporting in interviews typically take one of two forms. The simplest entails a concerted effort on the part of the interviewer to secure respondents' commitment to the study, to assure them that their answers will not be traceable, and to provide reinforcement and feedback over the course of the interaction (Cannell et al., 1981). In some circumstances involving sensitive items, the interviewer may make use of audio computer-assisted self-interview software (ACASI), in which questions are presented via computer, and the respondent answers using the keyboard. In this way, potentially embarrassing information need not be divulged verbally to the interviewer. If the respondent believes the promise of anonymity and confidentiality, the likelihood of truthful responding is enhanced.

Interview Structure

Once the researcher has decided what the interview is to cover, the next question is how structured the interview is to be. The answer to this question depends on the extent of information the researcher possesses about the phenomenon to be studied, and features of the respondent sample. If no hypotheses have been generated, and no questions have been decided upon, then **unstructured interviewing** should be used. However, if based on theory or previous research a definite list of hypotheses about responses has been developed, then **structured interviewing** is appropriate. Unstructured interviews may focus on a specific topic, but do not have an *a priori* schedule of questions to collect specific types of information. Structured interviews are used to request information using a predetermined schedule of questions and to ensure that all questions are asked the same way and in the same order for all participants. **Semi-structured interviews** lie somewhere between completely unstructured and completely structured interviews.

Unstructured interviews. Unstructured (or exploratory) interviews are characterized by freely generated questions by interviewers and unconstrained responses by interviewees. It does not demand specific types of information be collected. In unstructured interviews, neither the phrasing of the questions nor the allowable types of responses are constrained. There is no "interview roadmap" (structure) to guide the interviewer, whose queries are dictated by the dyadic verbal exchanges between respondent and interviewer, and by the general subject matter of the research. Unstructured interviews may be thought of as dynamic conversations between interviewer and respondent. The less structured the interview, the greater demands on the interviewer's competence. It is the responsibility of the interviewer to pursue leads that appear promising or informative. The number of such leads that ultimately prove productive usually is a function of the technical skill of the interviewer, the extent to which the interviewer is versed in the topic under study, and the rapport that has been developed during the interview.

Semi-structured interviews. In semi-structured interviews (the structured-nonscheduled interview), a set of predetermined topics are to be covered, but it is up to the interviewer to decide how to ask the questions, which usually are tailored to each interviewee. This interview form is a compromise between the completely unstructured exploratory interview and the completely structured interview to be discussed next. The semi-structured interview imposes on the researcher the necessity of obtaining certain specified types of information (hence the term, structured), but does not have an interview schedule specifying the way the information is to be obtained.

Suppose an interviewer wishes to obtain information about people's perceptions of same-sex marriage using a semi-structured interview. During the interview, given that there is no schedule that dictates the way this issue is approached, one respondent might be asked "What are your feelings about same-sex marriage?"; a second might be asked "What are your opinions about people of the same sex getting married?"; and a third interviewee may be asked "Should people of the same sex be allowed to marry?" These questions represent different ways of requesting the same structured information about the topic of same-sex marriage. The use of this technique is predicated on the assumption that some rough theoretical position has indicated the types of information desired for study purposes. In a semi-structured interview, the task of the interviewer involves requesting specific types of information as decided prior to the interview. The schedule of questions employed in completing the task is not.

As with the unstructured interview, the semi-structured interview is not used to best advantage in testing hypotheses, because the phrasing of questions varies from respondent to respondent, given that no fixed set of interview questions constrains the conversation. Semi-structured interviews require interviewers of considerable technical competence because they must be able to guide and redirect respondents to the critical information, but there is no schedule of prespecified questions to assist them in this task. Because the interviewer's work will consist of investigating prespecified types of information, he or she need not be overly sensitive to variation in responses that occur outside the topic boundaries initially established. This "narrow focus" relative to an unstructured interview can result in a loss of information, because the prespecified focus of the semi-structured interview constricts the research field. Although this constriction may result in a loss of wider knowledge about the topic, such techniques provide for a more intensive and focused investigation of the concepts and ideas thought to be most promising.

This is not to suggest that the scope of semi-structured interviews must be narrowly focused. The ground-breaking research program of Kinsey and his associates (1948) was accomplished primarily through semi-structured interviews. Kinsey understood that questions focused on various sexual practices would affect different respondents in different ways, and thus a standardized schedule was not indicated. The very sequence in which certain topics were introduced was varied according to features, such as age, education, socioeconomic status, etc. These interviews were structured insofar as interviewers were required to inquire about specific sexual habits, but they lacked a fixed schedule requiring a prescribed order of questioning, or specific question wording. Had Kinsey's interviewers followed a prespecified interview schedule, they almost certainly would have alienated a portion of the respondent sample. The use of the semi-structured interview form is especially appropriate when the sample is diverse, and thus phrasing of the questions can be tailored to everyone. However, if the sample is characterized by a more homogeneous respondent group, or the topics are anticipated to affect all respondents in the same way, then use of a predetermined schedule for the interview is recommended.

Structured Interviews. In a structured interview, the interviewer requests specific types of information, by asking the same list of questions in the same order for all interviewees, using a fixed schedule of questions. Considered the most systematic form of interview, the researcher

is discouraged from deviating from the standardized script. Even if the participant provides an interesting response that might shed light on the topic, the interviewer is not allowed to pursue it, as the schedule of questioning must be the same for all respondents. Any change of wording of the question is not permitted, as it might spur different responses from different participants. For this reason, structured interviews may be more challenging to conduct when participants are demographically diverse. The level of competence necessary to conduct a structured interview is less than that required for unstructured and semi-structured interviews, and this fact is reflected in interview cost savings (Schuman & Kalton, 1985). However, although the structured interview restricts deviations from the question script, it still calls for technical expertise, because a poorly trained or unmotivated interviewer can sabotage even the most highly structured research project.

Before proceeding with a detailed discussion of structured interviews, it is important to note a distinction between types of question constraint. Interviews may use **closed-ended questions**, in which the interviewer imposes a constraint on the possible options for responding. Interviews making use of this type of question are analogous to verbal multiple-choice tests – information is asked of a respondent, who is given a set of allowable answers from which to choose, e.g., “Are you a Catholic, Protestant, Jew, or Muslim?” If the possible range of options is not carefully considered, some (e.g., Buddhists and atheists) would be unable to choose among the options. An open-ended question without responses being constrained would be, “What is your religion?” allowing respondents freely to indicate any religion. Both questions seek the same information, but in the latter case, no constraint is placed on the allowable response, which must be coded after the interview.

The most obvious administrative difference between these approaches is that using **open-ended questions** can place greater demands on the interviewer, who must note the respondent’s replies. At the analysis phase, more costs can be incurred, especially on questions allowing for wide-ranging replies, because with this form of question, a system for classifying respondents’ answers must be developed before analysis can proceed (see Chapter 14 on the use of content analysis techniques for this purpose). With the closed-ended question format, the classification scheme is provided in advance. This simplifies the interviewer’s job, in addition to reducing analysis costs. But there is more to it than this benefit.

The minor administrative differences between open- and closed-ended questions are associated with major differences in the types of issues that interviews making use of these two question forms typically are designed to address. The structured interview making use of closed-ended questions allows for the greatest degree of standardization in the interviewer-respondent interaction. As in an experiment, all respondents are provided with a standard set of interviewer-produced stimuli (questions), along with a set of allowable answers. Interview schedules involving the same list of standardized questions are used in situations in which the investigator possesses considerable prior knowledge about the phenomenon of interest, and the options are constrained. Given that such question types are systematized by having the interviewer phrase the same questions in the same order, this method provides greater control and generates a more standard research context.

The required standardization of items and allowable responses characteristic of the structured-scheduled closed-ended interview can be accomplished only when the researcher’s specification of a participant’s potential range of responses is nearly complete. If the researcher unwisely uses questions constrained by only a few options without adequately defining the universe of potential responses, the item may provide unusable data. Interview items that give rise to very different reactions (sometimes called *self-stimulation effects*) can have a negative impact on standardization. If a question elicits heightened interest from 10% of the sample, for example,

problems regarding the comparability of these data with that of the other 90% must be considered, unless there is some way of differentiating those who were interested in the question from those who were not. The solution to this problem lies in the proper specification of question phrasing, consideration of potential responses, and their anticipated effect on respondents. Such systematization to derive a set of standardized interview questions can be assured only after considerable preliminary research.

Self-stimulation effects are not the only problem that premature use of closed-ended questions can produce. Even more likely are problems brought about by answers that force the interview to abandon the preset schedule. To return to our earlier example, suppose that in reply to the question, "Are you Catholic, Protestant, Jew, or Muslim?" our respondent were to respond "No." At this point, the closed schedule must be abandoned, because that response does not fit into the standardized scheme, and there is a good possibility that later questions, if contingent upon the response, will be inappropriate. Before proceeding in this circumstance, therefore, the interviewer must seek clarification of the respondent's answer. If the respondent meant, "No, I am none of these, I am an atheist," then later questions regarding the intensity of one's religious convictions, attendance at church, synagogue, or mosque, etc., would be inappropriate. If, however, the respondent's reply upon further questioning were found to mean, "No, I am none of these – I am a Buddhist," then the questions that follow on the closed-ended question may be usable. The clarification process forced upon the interviewer at this juncture, however, destroys the comparability of this interview with that of the others for whom the closed format was adequate. Use of a closed interview schedule that is insufficient to capture the respondents' range of possible or likely replies can destroy the utility of the interview.

Another potential disadvantage of using closed-ended questions is that they do not allow the interviewer to respond to novel information that might arise, but that is not a part of the prescribed schedule; the interviewer must "stick" to the script as it was written and cannot react to new information given by the respondent unless this can be done within the context of the follow-on questions. In such circumstances, interviewees often assume that the researcher is paying little heed to their replies, and under these conditions, maintenance of rapport becomes difficult. Despite these shortcomings, the structured-scheduled interview using closed-ended questions is most likely to be employed in hypothesis-testing contexts, because of the problems of noncomparability inherent in all the other types of interview formats. Given the demanding requirements that must be met before the closed-ended question can be employed, it is appropriate to ask whether the ideal of a completely comparable interview situation for all respondents can ever be attained in practice. The answer depends principally on the issue or issues on which the interview is focused.

Conducting the Interview

We have described the unstructured interview, the semi-structured interview, and the structured interview as if the operations that characterized these techniques were mutually exclusive. In practice, there are elements of all three types in almost all research interviews. Like most ideals, the descriptions presented are rarely realized, primarily because of the inevitability of uncontrolled, informal behaviors of the interviewer that occur during the administration of the schedule of questions. Previous research (Marquis & Cannell, 1969) using tape recordings of interviewers' behaviors demonstrated wide variations in language, even when completely structured interviews with fixed schedules were used – and the most experienced interviewers appeared to be the major culprits. A videotaped recording of these same interviews undoubtedly would have revealed even greater variation among interviewers, whose changes in seating

position and posture, smiles, nods of the head, changes in expression, and vocal intonation would also be readily apparent to respondents. And although the effects of these “paralinguistic” behaviors are not completely understood, they can affect an interview’s tone and progress.

The correspondence between the ideal, “pure interview forms” presented earlier and those arrived at in practice is a function of several variables, and we will consider some of the most important of these. However, it should be stressed that the apparent ubiquity of extraneous interviewer behaviors does not necessarily compromise the validity of the interview as a research tool. The so-called “extraneous” interviewer behaviors that are so apparent in any close inspection of almost any interview at times prove to be necessary adjuncts of this research technique. If an interviewer were to ignore the replies and nonverbal cues of the respondent, the sensitive person-to-person interaction that plays an important role in any interview would be forfeited and the quality of the obtained data adversely affected. “Rapport,” “interviewer-respondent interaction,” “respondent cooperation” are words and phrases emphasized in almost every interviewing handbook, but there are only a few things known of the ways and means of generating these states conducive to the interview process.

Obtaining Entry

An important task that should precede interview administration concerns the establishment of entry, or gaining permission to approach an individual or group of individuals for research purposes. Richardson et al. (1965) distinguished two qualitatively different types of entry situations – those in which a population of potential respondents is insulated from the interviewer by a gatekeeper, and those in which access to respondents is not controlled by a gatekeeper. Different approach strategies should be used in these different contexts.

A *gatekeeper* is some type of authority figure who controls access to potential respondents and who can affect the likelihood of respondents cooperating with the interviewer. Trying to skirt the gatekeeper can cause real problems, and thus should be avoided. Those who doubt this should imagine the consequences they would experience if they attempted to interview elementary school children during their afternoon recess or play period without first having secured the gatekeeper’s (i.e., the school authority, such as the school principal’s) permission. In this situation, it is quite possible that if permission was not obtained beforehand, the only information one would gain would be an indication of the speed with which the police responded to the call of an agitated educator. Examples of populations secured by a gatekeeper include elementary and high school students, members of unions, fraternities and sororities, athletic teams, adolescent gangs, church attendees, and hip hop groups.

The most obvious and direct strategy in gaining access to a protected group is to approach the person in control and state the aims and methods of the proposed research in a way that is understandable, nonthreatening, and accurate. It also is important to provide some rationale as to how participation can benefit the target group. This advice is complicated because the effective gatekeeper in many settings is not immediately obvious. For this reason, Richardson et al. (1965) suggest that the interviewer not press for an immediate entry decision. If the researcher has misidentified the gatekeeper (e.g., asking permission from a teacher when the school principal should have been contacted), but nevertheless convinced this person of the importance of the research, it is possible that the “misidentified” individual might intercede on behalf of the research. Forcing a pseudo-gatekeeper into a premature decision more often than not results in an outright rejection or, in the case in which entry is (apparently) secured, a later reversal by the real gatekeeper. It might appear more difficult to study samples protected by a gatekeeper than to investigate individuals not shielded in this manner. This is not necessarily true; the

gatekeeper's approval can legitimize the survey, and encourage respondents' cooperation. In populations that have no gatekeeper, this potentially facilitative influence is not available.

With respondent populations that are not protected by a gatekeeper, the researcher must approach each potential respondent directly and secure permission to be interviewed. Two techniques that have been shown to facilitate entry in these cases are letters of introduction and the "foot-in-the-door" approach. There is some evidence to suggest that a prior letter that alerts the potential respondent that he or she will be contacted later to participate in a research survey helps to increase the rate of participation (Brunner & Carroll, 1969; Cartwright & Tucker, 1969). The foot-in-the-door technique is based on the idea that securing a person's cooperation with a very minor request, and reinforcing this cooperation, facilitates their later cooperation with a major one (Freedman & Fraser, 1966). In applying this finding, Groves and Magilavy (1981) asked respondents to participate in a very minimal survey (two questions) and informed them that they might be called later to participate in a larger survey. Those who participated in the minimal survey were significantly more willing to cooperate with the later request than those who had not. Other researchers (e.g., Bloom et al., 2006; Souchet & Girandola, 2013) have used variations of the foot-in-the-door approach in difficult field contexts with good success.

After Entry

Once a potential respondent has been contacted directly or through a gatekeeper, several steps can be taken to ensure the individual's willingness to take part and cooperate with the interviewer.

Introduction of purpose. Assuming that the interviewer can at least get a foot in the door, there are a number of introductory procedures that seem to have a positive influence on the likelihood that a person will agree to cooperate. Cannell and Kahn (1968) recommend that the interviewer first provide a general description of the research project, then discuss the more specific research objectives, and finally outline the means of attaining these goals. Certainly the language the interviewer uses here will not be technical or scientifically rigid, nor is it likely that the initial introduction will be as extensive as that presented when seeking a gatekeeper's approval. Satisfying respondents' curiosity about the study and their role in the research process is both important and necessary if real cooperation is to be secured.

Method of selection. Having informed the potential respondent of the general nature of the research, disclosing information about how people were selected for study should be provided. If a specific "protected" group is being studied, and the gatekeeper has granted approval, this fact should be made known. The reasons for selection of the group to which the respondent belongs also should be mentioned.

If respondents who are not buffered by a gatekeeper are studied, some information about the sampling procedures employed in their selection should be given. A treatise on sampling design is not being suggested for this purpose, but rather some general information about the selection process should be mentioned, if such techniques were used. This step is sometimes skipped in the introductory phase of the interview, and this omission is unfortunate because respondents who are unsure about why they were "singled out" for the interview sometimes are less than completely candid in their answers. Interviewers who are transparent in describing the recruitment and sampling process invite participants to reciprocate by offering more transparent responses.

Agency. Cannell and Kahn's (1968) third step in the introduction process consists of identifying the organization or agency under whose auspices the study is being conducted. This procedure is even more important today than when it was suggested originally. Many unethical

sales organizations have employed pseudo-interviews to gain access to potential customers, and the public is more and more aware of the fact that persons asking for “a moment of your time to gather information on a very important issue” often are salespersons masquerading as social scientists. The resulting skepticism adversely affects the good will of potential respondents toward legitimate investigators. If satisfactory credentials can be presented, however, this difficulty can be overcome.

Anonymity. Many people who are not trained in the methods of social research assume that all survey responses can be traced directly back to their source – they are unaware that most analyses employ response pooling procedures that aggregate answers over the entire sample (or subsample) to infer general trends. Accordingly, the confidentiality or anonymity of an interviewee’s responses should be guaranteed, if possible. It is profitable in such circumstances to avoid collecting identifying information – e.g., asking for the participant’s full name – and to explain how the data will be aggregated and analyzed.

A willingness on the part of the potential respondent to take part in the interview might become apparent before the introductory sequence has been complete. The process sketched here paints an unduly pessimistic picture, for individuals often are quite eager to be interviewed and need not be persuaded through a long and arduous introductory process. There are many rewards accruing to interview respondents – emotional satisfaction, being able to express their views on matters of personal importance, the pride of being chosen to participate in a scientific investigation, the possibility of having their responses affect policy – all these factors make life much easier for the interviewer. The question then arises, “If I have secured entry by step 2 of the introductory process, should I continue through the entire sequence?” We believe this should be done, because in addition to securing cooperation, the steps enhance rapport. Neglecting any of the suggested procedures can compromise interview quality.

Interviewer Characteristics

Many factors other than those already mentioned can influence the interviewer’s chances of gaining entry and establishing rapport with respondents. Some of these reside in the characteristics of the interviewer. Included among these factors are the interviewer’s physical appearance, dress, race, accent, apparent socioeconomic status, and ethnic heritage. Whenever possible, it is wise to match the obvious demographic characteristics of the interviewer with those of the expected respondent sample, and many research firms try to do this. A complete matching is rarely possible, but there usually are some salient aspects that should be attended to. Research suggests the importance of matching the race of the interviewer with that of the respondent (e.g., Cotter et al., 1982), especially with sensitive topics concerning racial issues. People seem more reluctant to voice racial dissatisfactions with interviewers who do not match their own racial group. Findings of this type have persisted over the years, especially with racially sensitive issues, and appear to hold for Black and White respondents alike (Hatchett & Schuman, 1975; Schuman & Hatchett, 1974).

In more long-term interactions, the researcher is well advised to attend closely to a match of other social characteristics that might be important to respondents. Although complete matching is rarely possible, there may be some salient features of the respondents’ lifestyles that are shared by the interviewer that could facilitate their interaction. A good example of this form of matching is provided in William Foote Whyte’s (1955) investigation of adolescent gang members in “Cornerville.” The openness of interviewee responses was facilitated greatly by Whyte’s extensive knowledge of baseball facts, and his bowling skill (!), two interests his respondents shared avidly. If Whyte had approached his respondents as a Harvard researcher whose interests

did not carry him beyond the walls of the sociology library, the classic *Street Corner Society* might never have been written.

In addition to a match on demographic and lifestyle characteristics, there are other personal characteristics of the interviewer that may influence the relationship between interviewer and respondent, including the interviewer’s enthusiasm for the research, their professionalism, and apparent interest and friendliness with the respondent. Many of these factors cannot be directly controlled but depend on the availability and selection of experienced and well-trained interview personnel.

Informal Tactics

In addition to characteristics of the interviewer themselves, there are a number of informal behaviors and tactics that can be used effectively to gain and retain interviewee cooperation.

Initial question sequence. In Chapter 12, we raised the possibility that question order could influence the responses obtained in an interview. How one answers an early question may have a powerful influence on how later questions are answered. In addition to these considerations, it is important to understand that early interview questions can play a role in establishing rapport. Accordingly, the least threatening, least demanding, most general, and most easily answered questions should be presented first. As Kinsey et al. (1948) suggested, this order may vary because of respondent characteristics. Later, once cooperation is assured and the confidence of the respondent in the integrity of the interviewer is established, more difficult, specific, and sensitive issues may be broached.

Leading questions. Most of the early manuals on interviewing technique sounded a common injunction against the use of “leading questions.” A **leading question** is phrased in a way to suggest the expected answer or the premise that must be accepted to answer the question (see Table 15.1). Considering the effects of experimenter expectancy (Chapter 6) reinforces the apparent wisdom of this warning. Richardson and colleagues (1965) suggest that the injunction needs some qualification, arguing that under some admittedly constrained research conditions, the leading question could prove a useful and nonbiasing feature of proper interviewer technique. First, however, we must distinguish between two categories of leading questions, which have been termed *expectations* and *premises*.

A leading question in expectation form is a query whose wording alerts the respondent to the answer expected by the interviewer: “You voted for Biden, didn’t you?” is an example of a leading question based on an expectation. A better way to phrase this question is, “For whom did you vote?” Even with the modification, this question still requires the premise that the person voted, a faulty assumption if the interviews were conducted using a sample that also included nonvoters. The premise form of leading question contains within it one or more assumptions

TABLE 15.1 An example of a leading question exchange

Speaker	Response
Interviewer	Do you feel you have biases against people who suffer from mental illness?
Respondent	Of course not.
Interviewer	So you support the mental health clinic being established in your neighborhood.
Respondent	Are you nuts? I wouldn't risk my children's welfare in that way.
Interviewer	How so?
Respondent	By having a bunch of crazy people running around the neighborhood day and night.

that must be accepted if the question is to be answered. “When did you stop beating your wife?” is a classic leading question based on a premise. It presupposes the premises that (1) the respondent was or is married, (2) that he or she did, at least once, beat his or her wife, and (3) that he or she has stopped doing so. The general injunction against leading questions of the premise variety concerns queries with unfounded premises. But there are times when the premises underlying such questions are well founded and so can prove useful. The wife-beating question, for example, would not appear particularly ill-advised if asked of a respondent with a history of arrests for domestic abuse.

It is difficult to phrase questions to be completely free of premises. At the very least, a premise made in all interviews is that the respondent will communicate using a language the interviewer understands. Only when the premise is a completely uninformed guess would there appear to be much cause for alarm, not so much because of the potential for biased responses, but rather because it could compromise the interviewer–respondent relationship. Expectations and premises can be combined in the same question, with each varying in degree of strength. Of central importance is the degree to which the expectation or premise is founded on the interviewer’s anticipation of the respondent’s likely answer. If an expectation is completely accurate (and very often, accuracy can almost be assumed, based on earlier answers obtained in the interview), then there is little danger of bias in making use of it. Leading questions based on information generated earlier in the interview often are used by researchers to maintain the attention of the respondent over the course of the interaction. Well-informed expectations and premises indicate that the interviewer has been attentive to the interviewee’s responses (e.g., “As you just mentioned having beaten your wife, how many times have you done so?”). Thus, in some instances, using informed leading questions might improve rapport and respondent cooperation.

Although the proper use of leading questions depends on the interviewer’s near certain knowledge of the respondent’s likely answer, they need not always be used in anticipating the *correct* reply. A tactic used by experienced interviewers that sometimes pays dividends consists in deliberately missing the point of the respondent’s remarks (e.g., if aware that the interviewee is unmarried, the interviewer might ask “How many times have you beaten your wife?”). This tactic can result in an extensive elaboration of an earlier position, assuming that the interviewer’s “misinterpretation” is outrageous enough, and that the respondent is motivated to correct it (the participant might respond with, “No, you don’t understand, I’m not married. And if I were, I would never beat my wife”). The danger here is that the misinterpretation is not wrong enough, or that the respondent is not sufficiently motivated to correct the errant impression. An apparent complete and total misunderstanding of an individual’s opinion, however, often is sufficiently motivating to elicit an elaboration, especially if the issues at hand are important to the respondent.

The use of leading questions involving expectations and premises presupposes that rapport has been established, that the interviewer has a very good idea of the respondent’s likely reaction, and that the question is either totally correct or blatantly incorrect. Deploying leading questions with partially correct premises or expectations might have an adverse effect on the interview. A final piece of advice has to do with the frequency of deliberately leading questions: Overuse of these devices demeans the intelligence and sophistication of respondents, and can cause premature withdrawal from the interview.

Direction. When requesting elaboration of information, should the interviewer force specific clarifications and amplifications using a directive approach, or be completely nondirective, thus enabling the respondent to offer the clarification and amplification? The question of the degree to which an interviewer should direct the answers of a respondent is primarily of relevance in nonstructured and semi-structured interviews, and the answer depends on the

aims of the research. If in response to an inquiry, a respondent replied, “Yes, I think property taxes are much too high in this city,” an interviewer using a direct approach might steer the respondent’s attention by mentioning that property taxes support public schools, to determine the interviewee’s views toward tax reductions on these institutions considering the earlier expressed opinion that taxes were too high. Or, using a nondirective approach, the interviewer might simply restate the respondent’s position by saying, “You said that taxes are too high,” with the expectation that the respondent would be forced to clarify the original statement in response to the interviewer’s declaration (note that the inquiry is presented in declarative, rather than interrogative, form). To a nondirective prod of this type, the interviewee often will elaborate and defend the earlier position. Whether, during this nondirective process, the respondent eventually mentions the specific issue of interest (e.g., tax reduction effects on public schools) cannot be answered. This must be considered when gauging the appropriate degree of direction used. Nondirective approaches appear most useful in unstructured interviews, as their general purpose is to keep the researcher from imposing his or her own views about what is important in the situation under investigation.

Informal interviewer behaviors. As any experienced interviewer will testify, not all respondents are responsive. Merely asking a question does not guarantee a complete, understandable, and unambiguous response. In some cases, the fortunate interviewer is paired with an articulate person who fits the role of the perfect respondent; more often, even with the most cooperative of respondents, some prodding is required. The informal behaviors used to elicit an amplification or clarification of an unsatisfactory response, although relatively under investigated, are of real importance. Moreover, they are a ubiquitous component of all interviews – they are a constant feature of interviewer behavior (see Bradburn et al., 1979; Cannell et al., 1981; Marquis & Cannell, 1969).

A common behavior (or, in this instance, *non-behavior*) used to elicit elaborated responses is *silence*. It is important to understand that an interview is a research context involving interpersonal communication, and thus is subject to the same informal norms as other, more mundane conversations. In most conversations, cessation of verbal behavior on the part of one individual is interpreted as a cue to the other to begin another verbal exchange. In this way, each communicator’s response (or silence) acts as a stimulus to the other. An interviewer can make use of this reciprocity norm of interpersonal communication in the following manner: If a respondent’s answer is inadequate, the interviewer can fail to react – i.e., fail to re-initiate the conversation. Noticing that the answer did not provoke a reaction by the interviewer, the interviewee often will elaborate or amplify the earlier response. Of course, this tactic can be overused to the discomfort of the respondent – silence is a central feature of “stress interviews” (Chapple, 1953) – but when used sparingly, interviewer silence or nonresponse can stimulate respondents to elaborate their answers, with little danger of biasing the obtained data.

Verbal and nonverbal reinforcement. A different tactic that sometimes is used involves reinforcement. Phrases such as “good,” “fine,” “interesting,” and the like sometimes are used to encourage interviewees to amplify or elaborate a response. By reacting to respondents in this way, the interviewer demonstrates interest in their answers. This can strengthen rapport and assure continued cooperation. Unfortunately, whereas these tactics are undeniably motivating for respondents, the question of *what* is being motivated is not completely obvious. The “verbal reinforcement” literature suggests that an individual will express an opinion consistent with those that have been reinforced earlier in the interview (Fennis & Stroebe, 2010; Olson & Kendrick, 2008). Thus, an interviewer’s verbal reinforcement of a respondent for providing a complete or elaborated answer might be interpreted by the respondent as a reinforcement of the *content* of the reply. This could seriously affect subsequent answers. More subtle, but equally effective

reinforcers made by the interviewer are smiles, nods of the head, “uh-huhs,” etc. Actions of this type encourage the respondent, but their biasing effects may offset their utility.

Group Interviews and Focus Groups

In a **group interview**, the interviewer verbally requests information from more than one interviewee at the same time, but they are discouraged from interacting with each other. Interaction generally involves only interviewer and one interviewee (at a time). Group interviews are commonly employed as part of the hiring practice in organizations when there are many applicants for the same position, but rarely in research, except in focus groups, discussed later. Ordinarily, using multiple respondents in the same interview is not good practice, and is rare. To whittle down the pool of applicants for a job, however, interviews may be conducted with several candidates at the same time. Depending on the structure of the interview, there may be a predetermined order in which a person answers a question, followed by the next person who responds to the same question, and so forth. In other variations, the interviewer poses a question, and only participants who choose to answer do so. All the approaches for conducting proper one-to-one interviews – level of structure, whether a fixed schedule is used, participants’ sampling, and question wording – also apply to conducting group interviews.

One form of group interview this is often used in the early stages of a research project is the **focus group**. A focus group involves a moderator who obtains information by encouraging verbal discussion and interaction about a focused topic among a group of participants. In the typical focus group study, participants are recruited based on some criterion (gender, occupation, interests) relevant to the purpose of the investigation (Breen, 2006; Koppelman & Bourjolly, 2001). To facilitate interaction, participants often are seated around a table facing one another, not the moderator, who serves as the arbitrator who facilitates the focused discussion. The conversations are audio- or videotaped for later transcription.

Unlike participants in group interviews, those in focus groups are encouraged to interact and exchange ideas with one another, to stimulate greater depth of discussion and insights about a topic. Observation of verbal interactions among group members might produce insights about group values and norms. The group dynamic is intended to spur responses that would not have been obtained if social interaction was absent. A focus group is characterized by three features (Krueger & Casey, 2009; Wyatt et al., 2008):

- 1 *Qualitative data.* Quantitative data usually are not collected in focus groups, as questions such as “What percentage of the time do you use a condom?” are more appropriate in a questionnaire or interview format. A better fit would involve the question, “Discuss some reasons why people might or might not use a condom.”
- 2 *Homogeneous participant groups.* Generally, respondents in focus groups are drawn from groups sharing a critical feature. Examples might involve groups of respondents who have tested out a new consumer product, patients suffering from the same medical condition, former drug users, doctors who work the late shift in hospitals, expert chess players, and the like. A recommendation is that recruited participants should not know each other beforehand, as preexisting relationships and alliances will affect the group dynamic.
- 3 *Discussion of a focused topic.* The “focus” in focus group refers to a common stimulus or unifying topic that promotes the discussion, conversation, dialogue, or debate. In a structured focus group, where a scheduled list of questions is to be addressed, the moderator asks participants to converse about a topic, and the moderator shifts the conversation to the next

topic after obtaining sufficient responses. In unstructured focus groups, a general topic is presented, and participants are asked to converse freely about any aspect related to that topic. Here, participants dynamically drive the direction of the discussion topic.

Responses obtained in focus groups emerge not only from the interaction between the moderator and participants, but also from participants' interactions. Focus groups' norms and dynamics are dictated by the composition of their members, which influences the responses expressed by group members. This inherent feature is especially advantageous. It facilitates dialogue and disclosure on topics that require social support from similar others. For example, consider a focus group of adults who were formerly victims of child abuse. If one participant has the courage to share his or her experiences openly, this might foster greater willingness from others to self-disclose sensitive information. On the other hand, the group setting also might prove an obstacle for collecting data. For example, if the vast majority of group members openly express why they are pro-choice, the sole member who is not espousing this position may be susceptible to conformity pressure. To minimize this problem, the moderator should make it clear when setting the guidelines of discussion etiquette that the group is a safe environment in which everyone should express their opinions freely.

The moderator plays an important role in facilitating the conversation among participants. More vocal group members may monopolize the discussion, offering their opinions more frequently, while the quiet members may opt to contribute less or not at all. It is the moderator's responsibility to notice this and ask for the opinions of those who would rather be wallflowers ("How about you, what do you think?"). A good moderator requires the diplomatic skills to ensure that a single person does not "hog" the spotlight, or the data collected will be skewed toward the opinions of this member. It is challenging for the moderator to be sensitive to the responses of all participants and make sure all opinions are heard. For this reason, most focus group investigations involve a relatively manageable number (typically 6 to 12) of participants. An effective moderator should be highly trained on the topic to be able competently to address questions arising about the topic at hand. Also imperative are conflict resolution skills, so the moderator knows how to defuse tense situations, for example, when a member might constantly interrupt or make disparaging remarks about others' views.

Some focus groups have more than one moderator. An approach involving two moderators might be used if the moderators are trained to express opposite view points to help stimulate participant discussion of the pros and cons of an issue. This approach also may encourage participants to contemplate the entire range of opinions surrounding an issue.

Not all focus group studies use a moderator. In these highly unstructured focus groups, after giving instructions regarding rules for engaging in the discussion and information about the topic, the moderator leaves the room for a given period. The hope is that the absence of the moderator offers participants greater autonomy in openly expressing their uncensored views. Discussions lacking a moderator may be monitored, but because unmoderated focus groups require the questionable assumptions that participants will be self-motivated to engage each other in conversation, not stray from the main topic, and cover all the topics within the allotted time, groups lacking a moderator are rarely conducted and generally not recommended.

For research purposes, focus groups usually are used as pilot studies in the formative stages in a new line of research or to obtain preliminary evidence to learn about an understudied topic. The findings obtained in focus groups might be used to gain insights for a subsequent quantitative study. In scale development and validation research, focus groups are used to establish the content validity of a construct.

Conclusion

As we have shown, the interview method can take many forms, and can be applied to a range of disparate issues. Its general utility is widely appreciated by researchers, which is reflected in its widespread use. The method is flexible and can be applied at almost any stage of the research process. Interviews usually are administered face-to-face or on the telephone, and are used to study people's reconstruction of past events or current beliefs or behavioral reports. Different interview structures are determined by the needs of the research, and the planned uses of the resulting data. The chapter that follows will present information on the ways in which the verbal responses collected in an interview study can be distilled into useful scientific data.

Notes

- 1 Interviews, of course, are used in many contexts other than basic research data collection. The clinical interview is a valuable tool of diagnosis and treatment in mental health settings, and extensive participant interviews often accompany the debriefing phase of a laboratory experiment. However, these specialized uses of the in-depth interview technique are beyond the purview of this text. In the present chapter, we will focus on the use of the interview in the general context of survey research.
- 2 Of course, these findings presume that the respondent can be reached by phone in the first place. Given that approximately 95% of American households have telephones, this form of contact has not been problematic, at least in industrialized nations. However, in recent years the increasing use of answering machines, voice mail, and caller-id to screen out incoming phone calls has made access to potential participants for phone interviews somewhat more difficult (see discussion of sampling issues in Chapter 11).
- 3 Those who would like an overview of commonly employed measures (to use in developing their own specific item set) would be well-advised to consult Van Dusen and Zill (1975), and the report of the American Psychological Association's (APA) task force on socioeconomic status (APA, 2007).
- 4 These factors probably interact with the specific event outcome in determining accuracy of recall. We know that pleasant outcomes are more likely to be remembered, especially those that have occurred recently; thus, winning \$5,000 in the lottery is probably more readily recalled than losing \$500 in a poker game, not only because it is a more pleasant outcome, but also because it has a more enduring impact (Cannell & Fowler, 1963; Turner, 1982).
- 5 It is important to note that the principles that govern the reliable (and valid) measurement of attitudes apply equally to the measurement of behavioral intentions.
- 6 If accuracy of response is absolutely essential, it is important that the respondent know that choosing the "don't know" option is not tantamount to an admission of ignorance.

References

- American Psychological Association Task Force on Socioeconomic Status (2007). *Report of the APA task force on socioeconomic status*. Washington, DC: American Psychological Association.
- Andrews, P. W., Gangestad, S. W., Miller, G. F., Haselton, M. G., Thornhill, R., & Neale, M. C. (2008). Sex differences in detecting sexual infidelity: Results of a maximum likelihood method for analyzing the sensitivity of sex differences to underreporting. *Human Nature, 19*, 347–373.
- Bartholomew, K., Henderson, A. J. Z., & Marcia, J. (2000). Coding semistructured interviews in social psychological research. In H. Reis & C. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 286–312). Cambridge: Cambridge University Press.
- Bergen, G., Shults, R. A., Beck, L. F., & Qayad, M. (2012). Self-reported alcohol-impaired driving in the U.S., 2006 and 2008. *American Journal of Preventive Medicine, 42*, 142–149.
- Bessa, M. A., Mitsuhiro, S. S., Chalem, E., Barros, M. M., Guinsburg, R., & Laranjeira, R. (2010). Underreporting of use of cocaine and marijuana during the third trimester of gestation among pregnant adolescents. *Addictive Behaviors, 35*, 266–269.
- Bishop, G. R., Oldendick, R. W., Tuchfarber, A. J., & Bennett, S. E. (1980). Pseudo-opinions on public affairs. *Public Opinion Quarterly, 44*, 198–209.

- Bloom, P. N., McBride, C. M., Pollak, K. I., Schwartz-Bloom, R. D., & Lipkus, I. M. (2006). Recruiting teen smokers in shopping malls to a smoking-cessation program using the foot-in-the-door technique. *Journal of Applied Social Psychology*, 36, 1129–1144.
- Burnburn, N. M., Sudman, S., Blair, E., Locander, W., Miles, C., Singer, E., & Stocking, C. (1979). *Improving interview method and questionnaire design*. San Francisco: Jossey-Bass.
- Breen (2006). A practical guide to focus-group research. *Journal of Geography in Higher Education*, 30, 463–475.
- Brunner, G. A., & Carroll, S. J., Jr. (1969). The effect of prior notification on the refusal rate in fixed address surveys. *Journal of Advertising*, 9, 440–453.
- Cain, P. S., & Treiman, D. J. (1981). The dictionary of occupational titles as a source of occupational data. *American Sociological Review*, 46, 253–278.
- Cannell, C. F., & Fowler, F. J. (1963). Comparison of a self-enumerative procedure and a personal interview: A validity study. *Public Opinion Quarterly*, 27, 250–264.
- Cannell, C. F., & Kahn, R. L. (1968). Interviewing. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (2nd ed., Vol. 2, pp. 526–595). Reading, MA: Addison-Wesley.
- Cannell, C. F., Miller, P. V., & Oksenberg, L. (1981). Research on interviewing techniques. In S. Leinhardt (Ed.), *Sociological methodology*. San Francisco: Jossey-Bass.
- Cartwright, A., & Tucker, W. (1969). An experiment with an advance letter on an interview inquiry. *British Journal of Preventive Social Medicine*, 23, 241–243.
- Chapple, E. D. (1953). The standard experimental (stress) interview as used in interaction chronograph investigations. *Human Organization*, 12, 23–32.
- Cotter, P. R., Cohen, J., & Coulter, P. B. (1982). Race-of-interviewer effects in telephone interviews. *Public Opinion Quarterly*, 46, 278–284.
- Denis, C., Fatséas, M., Beltran, V., Bonnet, C., Picard, S., Combourieu, I., ... & Auriacombe, M. (2012). Validity of the self-reported drug use section of the addiction severity index and associated factors used under naturalistic conditions. *Substance Use & Misuse*, 47, 356–363.
- Dijkstra, W., & van der Zouwen, J., (Eds.) (1982). *Response behavior in the survey-interview*. New York: Academic Press.
- Dillman, D. A. (1978). *Mail and telephone surveys*. New York: Wiley.
- Fennis, B. M., & Stroebe, W. (2010). *The psychology of advertising*. New York: Psychology Press.
- Freedman, J. L., & Fraser, S. C. (1966). Compliance without pressure: The foot-in-the-door technique. *Journal of Personality and Social Psychology*, 4, 195–202.
- Groves, R. M., & Kahn, R. L. (1979). *Surveys by telephone*. New York: Academic Press.
- Groves, R. M., & Magilavy, L. J. (1981). Increasing response rates to telephone surveys: A door in the face for the foot-in-the-door? *Public Opinion Quarterly*, 45, 346–358.
- Hajebi, A., Motevalian, A., Amin-Esmaili, M., Hefazi, M., Radgoodarzi, R., Rahimi-Movaghar, A., & Sharifi, V. (2012). Telephone versus face-to-face administration of the structured clinical interview for diagnostic and statistical manual of mental disorders, fourth edition, for diagnosis of psychotic disorders. *Comprehensive Psychiatry*, 53, 579–583.
- Hatchett, S., & Schuman, H. (1975). White respondents and the race-of-interviewer effect. *Public Opinion Quarterly*, 39, 523–528.
- Herman, J. B. (1977). Mixed-mode data collection: Telephone and personal interviewing. *Journal of Applied Psychology*, 62, 399–404.
- Hodge, R. W., Siegel, P. M., & Rossi, P. H. (1964). Occupational prestige in the United States. *American Journal of Sociology*, 70, 286–302.
- Hollingshead, A. de B. (1949). *Elmstown's youth, the impact of social classes on adolescents*. New York: Wiley.
- Hong, Y., Fang, X., Zhou, Y., Zhao, R., & Li, X. (2011). Factors associated with sexually transmitted infection underreporting among female sex workers in China. *Journal of Women's Health*, 20(1), 129–136.
- Hornik, R., & Orwin, R. (2012). Robert Hornik and Robert Orwin on Stephen Magura's "failure of intervention or failure of evaluation: A meta-evaluation of the national youth anti-drug media campaign evaluation. *Substance Use & Misuse*, 47, 1436–1438.
- Kinsey, A. C., Pomeroy, W. B., & Martin, C. E. (1948). *Sexual behavior in the human male*. Philadelphia: Saunders.
- Cluegel, J. R., & Smith, E. R. (1982). Whites' beliefs about blacks' opportunity. *American Sociological Review*, 47, 518–532.

- Koppelman, N. F. & Bourjolly, J. N. (2001). Conducting focus groups with women with severe psychiatric disabilities: A methodological overview. *Psychiatric Rehabilitation Journal*, 25, 142–151.
- Kroner, D. G., Mills, J. F., & Morgan, R. D. (2007). Underreporting of crime-related content and the prediction of criminal recidivism among violent offenders. *Psychological Services*, 4, 85–95.
- Krosnick, J. A., & Fabrigar, L. R. (1997). Designing rating scales for effective measurement in surveys. In L. E. Lyberg, P. Biemer, M. Collins, & E. D. De Leeuw (Eds.), *Survey measurement and process quality* (pp. 141–164). New York: Wiley.
- Krosnick, J. A., & Fabrigar, L. R. (2001). *Designing great questionnaires: Insights from psychology*. Oxford: Oxford University Press.
- Krosnick, J. A., Holbrook, A., Berent, M. K., Carson, R. T., Hanemann, W. M., Kopp, R. J., Mitchell, R. C., ... & Conaway, M. (2002). The impact of “no opinion” response options on data quality: Non-attitude reduction or an invitation to satisfice? *Public Opinion Quarterly*, 66, 371–403.
- Krueger, R. A., & Casey, M. A. (2009). *Focus groups: A practical guide for applied research* (4th ed.). Thousand Oaks, CA: Sage.
- Lee, S., Tsang, A., Lau, L., Mak, A., Ng, K. L., & Chan, D. M-L. (2008). Concordance between telephone survey classification and face-to-face structured clinical interview in the diagnosis of generalized anxiety disorder in Hong Kong. *Journal of Anxiety Disorders*, 22, 1403–1411.
- Magura, S. (2010). Validating self-reports of illegal drug use to evaluate national drug control policy: A reanalysis and critique. *Evaluation and Program Planning*, 33(3), 234–237.
- Marquis, K. H., & Cannell, C. F. (1969). *A study of interviewer-respondent interaction in the urban employment survey*. Ann Arbor: Survey Research Center, University of Michigan.
- Olson, M. A., & Kendrick, R. V. (2008). Origins of attitudes. In W. D. Crano & R. Prislin (Eds.), *Attitudes and attitude change* (pp. 111–130). New York: Psychology Press.
- Parry, H. J., & Crossley, H. M. (1950). Validity of responses to survey questions. *Public Opinion Quarterly*, 14, 61–80.
- Quinn, R. P., Gutek, B. A., & Walsh, J. T. (1980). Telephone interviewing: A reappraisal and a field experiment. *Basic and Applied Social Psychology*, 1, 127–153.
- Ray, M. L., & Webb, E. J. (1966). Speech duration effects in the Kennedy news conferences. *Science*, 153, 899–901.
- Richardson, S. A., Dohrenwend, B. S., & Klein, D. (1965). *Interviewing: Its forms and functions*. New York: Basic Books.
- Richter, L., & Johnson, P. B. (2001). Current methods of assessing substance use: A review of strengths, problems, and developments. *Journal of Drug Issues*, 31, 809–832.
- Rossi, P. H., Wright, J. D., & Anderson, A. B. (Eds.) (1983). *Handbook of survey research*. New York: Academic Press.
- Rugg, D. (1941). Experiments in wording questions: II. *Public Opinion Quarterly*, 5, 91–92.
- Schaps, E., & Sanders, C. (1970). Purposes, patterns, and protection in a campus drug-using community. *Journal of Health and Social Behavior*, 11, 135–145.
- Schuman, H., & Hatchett, S. (1974). *Black racial attitudes: Trends and complexities*. Ann Arbor, MI: Institute for Social Research, University of Michigan.
- Schuman, H., & Kalton, G. (1985). Survey methods. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology*. Reading, MA: Addison-Wesley.
- Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. New York: Academic Press.
- Smith, E. R., & Kluegel, J. R. (1984). Beliefs and attitudes about women’s opportunity. *Social Psychology Quarterly*, 47, 81–95.
- Smith, G. T., McCarthy, D. M., & Goldman, M. S. (1995). Self-reported drinking and alcohol-related problems among early adolescents: Dimensionality and validity over 24 months. *Journal of Studies on Alcohol*, 56, 383–394.
- Souchet, L., & Girandola, F. (2013). Double foot-in-the-door, social representations, and environment: Application for energy savings. *Journal of Applied Social Psychology*, 43, 306–315.
- Sudman, S. (1980). Reducing response errors in surveys. *Statistician*, 29, 237–273.
- Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco: Jossey-Bass.

- Turner, C. F., & Martin, E. (Eds.) (1982). *Surveys of subjective phenomena*. Cambridge, MA: Harvard University Press.
- Van Dusen, R. A., & Zill, N. (Eds.) (1975). *Basic background items for U.S. household surveys*. Washington, DC: Center for Coordination of Research on Social Indicators, Social Science Research Council.
- Visscher, T. L. S., Viet, A. L., Kroesbergen, H. T., & Seidell, J. C. (2006). Underreporting of BMI in adults and its effect on obesity prevalence estimations in the period 1998 to 2001. *Obesity, 14*, 2054–2063.
- Weaver, C. N., & Swanson, C. L. (1974). Validity of reported date of birth, salary, and seniority. *Public Opinion Quarterly, 38*, 69–80.
- Whyte, W. F. (1955). *Street corner society*. Chicago: University of Chicago Press.
- Wyatt, T. H., Krauskopf, P. B., & Davidson, R. (2008). Using focus groups for program planning and evaluation. *Journal of School Nursing, 24*, 71–77.

16

CONSTRUCTION OF QUESTIONNAIRES AND RATING SCALES

Studying people's beliefs, attitudes, values, and personalities is a central research preoccupation of the social sciences. Typically, we use questionnaires or scales to measure these internal states or dispositions. Such measures rely on **self-report**, by asking respondents to provide answers to a set of questions or scale items that inquire about their personal thoughts, feelings, or behaviors. Using these measures requires that we consider variations in responses among respondents as meaningful, and not attributable to mere measurement error. In theory, all participants are expected to interpret the questions or items used in a scale or questionnaire identically. That is, a given questionnaire or scale item is assumed to mean the same thing to all participants, and differences in their responses to these hypothetically identical items are assumed to reflect real differences in their underlying dispositions.

Questionnaires

Two complementary approaches characterize attempts to establish measures to assess differences among people in their internal states or dispositions. For convenience, we will term the slightly less formal measures *questionnaires*, and the more rigorously designed measures *rating scales*. A **questionnaire** involves a single item or two to assess each construct, and typically is used because participants are unwilling, unable, or unlikely to take part in a longer assessment, or because the issue is one that admits to few options. Nationally representative polls of voter intentions (Candidate A or B?) or quality-of-life ratings often use questionnaires. After all, there are not many ways to ask, "For whom will you vote in the coming election?" **Rating scales** involve multiple items (to be meaningful, Cronbach's alpha should have a minimum of three items – see Chapter 3) to assess each construct. Rating scales typically are used by researchers with access to participants more willing to take part in a longer assessment. Questionnaire and rating scale construction are integral to many of the data collection methods discussed throughout this text, so it is important to have a good understanding of their strengths and weaknesses.

Some researchers do not have the luxury of access to participants for a long period of time and thus make use of questionnaires. That is, we are unable to use multiple items to tap a person's views of the same target – a person, event, or object. From earlier chapters, remember the benefit of assessing a target with multiple items reflects our acknowledgment that a single item

is insufficient to capture or triangulate a construct meaningfully. Sometimes, the limitation is imposed because of time costs: when conducting a national survey, for example, adding even a single item can be prohibitively expensive, so we must be content with an item or two to assess people's thoughts and feelings on a given topic, say, one's choice of President or one's position on gun control. In other cases, many concepts are under scrutiny, and to use multiple items to tap each of them would create an overly long instrument. Fears of participant fatigue and the accompanying loss of data quality motivate the use of a single item per concept. Some questionnaires are lengthy, measuring many concepts, but for our purposes, we will consider the more typical brief questionnaire.

Questionnaires may be administered in the context of an interview study (Chapter 15) or in written form or online as a self-administered measure. There are no formal rules for questionnaire design, but considerable wisdom has grown around their construction because of the intense focus on such measuring instruments over the past 70 years in social science (see Fabrigar et al., 2005; Krosnick & Fabrigar, 2001). We present some rules of thumb that are commonly followed. As will be seen, these rules will apply as well in the development of rating scales. All these recommendations apply whether the questionnaire is a written self-administered instrument, an online survey, or an interview.

Wording Questions

The first rule is to determine what you want to know. The more direct the phrasing of the question, the more likely is its true meaning to be understood. Complicated questions are more likely to be misunderstood or misread, and less likely to provide valid information. To capture information on binge drinking, the researcher should first define binge drinking to participants ("Binge drinking is consuming four or more alcoholic beverages in a single occasion"), then the question ought to be precisely phrased as "How many times, if any, have you binged on alcohol in the past (*week, month, year, or in your life*)?"

A second rule of thumb is to use short, simple sentences containing a single grammatical clause if possible when developing queries. This rule helps avoid the use of a **double-barreled question** – a single item that asks more than one question at once. Asking "Do you think civil rights activists have gone too far and that the government should crack down on militant civil rights organizations in this country?" is an example of a double-barreled question. The problem with such questions, of course, is that it is not easy to know which of the two queries contained in the item the respondent is answering. This compound sentence should be rewritten as two questions: "Do you think civil rights activists have gone too far?" and "Do you think the government should crack down on militant civil rights organizations in this country?"

Open- and Closed-Ended Questions

A distinct advantage of questionnaires over scales is the capacity to use items with an open-ended response format. An open-ended item is one that poses a question but does not constrain the answer. The advantage of open-ended questions (e.g., "Who is your favorite mayoral candidate?") over closed-ended ones (e.g., "Who is your favorite mayoral candidate: Smith, Jones, or Bradley?") is that the former does not force respondents to choose among a limited set of response options. If the respondent's answer to the questions posed here were "Johnson," the first would capture it, the second would not. Visser et al. (2000) have argued that open-ended questions are clearly preferable to closed-ended questions in questionnaire construction contexts. However, whereas open-ended questions may tap respondents' views with greater fidelity

than closed-ended items, their added sensitivity comes with a cost, which involves developing methods of analyzing the free responses that the open items generate. In the mayoral example, analyzing the content of open-ended answers would not prove onerous. However, respondents must be able to recall the name of their preferred choice, which may prove a disadvantage of an open-ended question. Without a set of response options, some voters might be unable to recall their preferred candidates but would be able to recognize and select that option if reminded with a closed-ended multiple-choice format or a ballot.

Whenever the list of possible options is relatively limited, and readily anticipated, coding answers becomes an almost mechanical task. So, for example, if we were to ask, “Whom do you believe is responsible for California’s energy problems?” informed respondents would most likely have a rather limited list of possibilities: avaricious electric companies, unscrupulous politicians, careless consumers, and a few other nefarious eco-villains. However, if the issue is one that admits to a host of possible answers, e.g., “What should the U.S. do about global warming?” a coding scheme is necessary, and its development will prove costly. Whether the cost is tolerable depends in part on available resources, the complexity of the issue, and the number of topics being studied. In contexts in which the questionnaire contains many issues, each of which requires constructing a coding scheme, the costs can be high. Sometimes, however, open-ended questions seem to be worth the trouble and complexity of analysis (Visser et al., 2000).

Question Ordering

In our discussion of interview methodology (Chapter 15), we stressed the importance of establishing rapport to help ensure the quality of the interviewer and interviewee interaction, and the truthfulness and completeness of respondents’ answers. In many self-report questionnaire studies, there is little, if any opportunity to establish rapport. The items are posed by a more or less anonymous questioner, with little attempt at developing a relationship with the participant, or are presented on a printed page with a brief introduction of purpose. In circumstances like these, question order may be crucial.

In questionnaire development, the analogue of the rapport-building process requires that the least threatening items be presented first. Only after the respondent has become comfortable with the research, and somewhat committed to it by virtue of answering several questions, should more personal or threatening questions be presented. For example, in research on adolescent drug use, nonthreatening queries are commonly presented before items assessing use of illegal or illicit substances are posed. This is the reason why demographic characteristics, which involve rather sensitive personal information (e.g., income, race/ethnicity), should be presented near the end of a questionnaire. Sometimes the ordering of items determines whether a respondent chooses to continue in the study, and this is not a trivial concern.

Another issue related to question order has to do with the possibility that one’s earlier answers may affect later ones. For example, suppose we were to ask, “Should freedom of speech be absolutely guaranteed in this country?”, most respondents would answer this question affirmatively. However, the affirmation of this fundamental human right might have ramifications on later answers having to do with the value of freedom – of thought, of expression, and of action. Affirming free speech probably inclines respondents to more liberal views on later items. However, suppose instead we were to ask, “Should hate speech be banned?” As before, a reasonable proportion of the respondent sample probably would answer this question affirmatively. In this case, later questions having to do with freedom of expression, thought, and action might be colored by this earlier response, but in a way opposite to that of the first example. The earlier items

might “prime” the mindsets of participants when they respond to subsequent items (Sudman & Bradburn, 1982).

If questionnaire items were presented to all respondents in identical order, the problems caused by this early item influence could bias the research outcome (LeCount, 2022; Magelssen et al., 2016; Yeung et al., 2020). To combat question order effects, some researchers recommend that questionnaire developers who fear the problem of influence from early items on later items counterbalance or randomize the order of items across participants (see Rasinski et al., 2012; Tourangeau & Rasinski, 1988; Tourangeau et al., 2000).¹ For self-administered written questionnaires, the same process applies. For questionnaires delivered by telephone interviewers, the question ordering can be varied using computer software preprogrammed to sequence the order of items. This reordering should not involve the entire questionnaire. The least threatening questions, as noted, should appear first. However, within the entire questionnaire, the order of items that might be reactive should be counterbalanced or randomized across blocks of participants. Thus, if a questionnaire developer anticipates that answers to any of the six items dealing with a particular topic might mutually influence the other answers, the order of those items might be counterbalanced or randomized across participants (Schuman & Presser, 1981; Standing & Shearson, 2010).

Dropout and the No Opinion Response Format

Researchers want to avoid losing respondents if possible. Whether respondent loss is attributable to a person’s refusal to initiate the questionnaire or to complete it once begun, the loss of respondents represents a threat to the generalizability of research results. Persuading potential respondents to take part in our research has been covered elsewhere (e.g., see Chapter 15). Here we are concerned with respondents who virtually drop out of the research by refusing to answer one or more questionnaire items. In these instances, many common statistical tests that apply listwise deletion will remove such respondents from the analysis.

A particularly difficult issue that affects the likelihood that a respondent will complete *all* items of a questionnaire is the inclusion or non-inclusion of a “no opinion” (or “don’t know”) option. Some questionnaires allow respondents to indicate that they hold no opinion on an item (or don’t know the answer to a question in a knowledge scale); others allow for a “neutral” response; still others do not provide a “middle of the road” category, forcing respondents to take a stand (albeit a weak stand) on one side or another on the issue, as no “middle” response option is offered, as in **forced choice** measures.

Investigations of the effects of these variations in response format have been conducted for many years (Converse, 1964; Rubin, 1987), and there are good arguments on both sides regarding whether to include a no response option. Respondents generally seem to prefer the no opinion (or don’t know) option (Beatty et al., 1998; Lee & Kanazawa, 2000; Luskin & Bullock, 2011), but Krosnick and colleagues (2002) suggest that including a “no response” option may encourage respondents to avoid the cognitive work involved in answering survey items, and thus discourage its use. Tourangeau et al. (2004) also warned against the use of the option, insofar as it may affect respondents’ interpretation of other available response options. In general, the advantages of allowing a middle option appear outweighed by the negative possibilities. In cases in which a good understanding of participants’ feelings about a given issue is particularly critical, it is advisable to provide measures that allow for a clear interpretation of the meaning of the nonresponse option, if it is used. Wegener et al. (1995) have suggested different measures that might help unravel the meaning of a “don’t know” response, and we direct the interested reader to this work.

Constructing Rating Scales

The approach to creating effective items for questionnaires applies to rating scale development as well. Rating scales are formalized versions of questionnaires. The difference is that whereas a single item usually is used to represent each concept in a questionnaire, multiple items are used in creating scales. Because scales use multiple items to triangulate on, or define a concept, they are more appropriately used to measure attitudes, values, or personality dispositions, reflecting the view that people's attitudes or beliefs are not singularly defined.

When assessing individual differences on attitude or personality scales, each item often takes the form of a statement the participant is asked to endorse or reject. Louis Thurstone developed one such scaling approach, and although his model has been supplanted by more modern approaches, it forms the basis of many widely used scaling procedures today.

Thurstone's Method of Equal-Appearing Intervals

Developing rating scales to measure beliefs, opinions, and attitudes represents an important aspect of the history of social science research in the last century (Crano & Lac, 2012). First attempted by Thurstone (1928, 1931; Thurstone & Chave, 1929), attitude assessment has become one of the social sciences' most important and persistent preoccupations. In the typical Thurstone scale (called "the method of equal-appearing intervals"), respondents are asked to endorse the scale item or items with which they agree. Items are designed so that a single item, or a highly restricted range of items, should be endorsed by the respondent, and those that are considerably more extreme or less extreme than the chosen alternative should be rejected. Items of this type have been termed *nonmonotone* (Coombs, 1950) or *noncumulative* (Stouffer, 1950), because it makes little sense to sum a respondent's scores over all items of the scale of this type. Agreement with one item does not imply an increased probability of agreement with any other item on the scale.

In practice, it is difficult to develop scales based on the method of equal-appearing intervals. Nunnally (1967, p. 69) has persuasively illustrated this point by asking,

...How could one find spelling words such that each would be correctly spelled only by persons in a narrow band of the attribute of spelling ability? An item that 'peaked' at the lower end of the scale would be one that is spelled correctly only by rather poor spellers. For an item that peaked in the middle...very few people with superior ability in spelling would give a correct response.

We encounter these difficulties when devising attitude scales based on Thurstone's model. Nonetheless, even though alternative models of scale construction are employed in most instances today, understanding Thurstone's approach is important as it forms the historical basis for many of today's more widespread alternatives.

A scale based on Thurston's method of equal-appearing intervals is developed and constructed in four phases. The first phase in the scale construction process requires the researcher to generate many *potential* items, all of which appear at least initially assess the construct, object, norm, or attribute of interest. Enough items should be developed to cover the complete range of possible evaluations of the critical object. Items should be concise and worded in such a way that their meaning is clear. Double-barreled items should be avoided, as should items on which either complete acceptance or complete rejection by most members of the respondent sample might be expected.

In the second phase of the scale construction process for the method of equal-appearing intervals, a group of judges that is part of the research team helps develop the psychometric properties of the scale. Each judge independently estimates the degree of favorability or unfavorability expressed by each item toward the critical attitude object. Traditionally, an 11-point response format is used, with the endpoints bounded by the numbered phrases “1 (extremely favorable)” and “11 (extremely unfavorable).” The judges are instructed to disregard their own personal attitudes during categorization of each using the 11-point judgment scale.

The third phase involves the selection of items for the final scale. Based on judges’ ratings, the investigator determines the mean favorability rating for each item and its standard deviation. An item with a large standard deviation is a danger signal, because it suggests that there is disagreement among judges about the favorability of that item. This result usually occurs because of item ambiguity, violating a central assumption that each item should be interpreted identically by all judges. Thus, items exhibiting high standard deviations on judges’ ratings should be discarded. From the pool of possibilities that remain, a reduced number of items are chosen to constitute the final version of the scale. Items are chosen so that the items derived from the judges’ ratings cover the entire range of possible evaluations of the attitude object. Items are chosen so that when they are arranged from lowest to highest order based on judges’ mean values, approximately equal differences in intervals are maintained between the judges’ means of successive items, as implied by the name of the method.² Although the term “content validity” was not coined when Thurstone developed the method of equal-appearing intervals, the three phases of this scale construction process were intended to establish content validity of the final set of items in capturing the entire conceptual bandwidth or range of the construct of interest.

Next, the researcher shifts from the *scale construction* phases to the *scale administration* phase. Administering the final set of items to participants is the fourth and final phase. Two forms of responding to Thurstone scales have been used for participants. In earlier research, respondents were asked to circle the item with which they most strongly agreed, and then one or two others with which they agreed but were not their favorite. This approach did not admit to ready analysis of the internal consistency reliability of the measure, and for this reason, the approach has been largely supplanted by other scaling methods.

In studies using Thurstone’s approach today, respondents are instructed to indicate whether they agree or disagree with each of the items. Other variations of this response format include answering “yes” or “no,” or “true” or “false” to each item. The average of the judges’ value of the items chosen by a respondent is taken as that individual’s attitude toward the object or issue. For example, consider the 12 items in Table 16.1, which illustrates use of Thurstone’s method of scaling approach.³ The mean values that judges previously assigned to each item are presented in the example. However, the item values would not be presented to participants – they are illustrated here for your use. In addition, instead of presenting the items in a systematic order, items of different favorability ratings would have been mixed randomly, as was done here. The mean scale value of the items each of the respondents had endorsed would be their assigned score on Thurstone’s attitude toward the church scale. For example, if a participant were asked to endorse three most favored items, the mathematical average of the values derived from judges’ previously assigned values corresponding to those items would be computed as that participant’s score on the construct.

Reliability. A major common indicator of scale quality, internal consistency (Chapter 3), is not a meaningful concept in the context of Thurstone scales that made use of the original scoring method (choose your favorite item), because such measures made use of participants’ overall responses to only a few items of a scale. In this case, we have information only on the items the respondent endorsed. However, test-retest methods of reliability estimation might be used to

TABLE 16.1 Twelve items from the more negative end of Thurstone and Chave's (1929) "Attitudes toward the church scale"

Item #	Item	Scale value
1	I think the teaching of the church is altogether too superficial to have much social significance.	8.3
2	I feel the church is petty, always quarreling over matters that have no interest or significance.	8.6
3	I respect any church-member's beliefs but I think it is all "bunk."	8.8
4	My experience is that the church is hopelessly out of date.	9.1
5	I think the church seeks to impose a lot of worn-out dogmas and medieval superstitions.	9.2
6	I think the church is hundreds of years behind the times and cannot make a dent on modern life.	9.5
7	I think the church is a hindrance to religion for it still depends on magic, superstition, and myth.	9.6
8	The church represents shallowness, hypocrisy, and prejudice.	10.4
9	I regard the church as a static, crystalized institution and as such it is unwholesome and detrimental to society and the individual.	10.5
10	I think the church would be better off if churches were closed and the ministers sent to some useful work.	10.5
11	I think the organized church is an enemy of science and truth.	10.7
12	I think the church is a parasite on society.	11.0

estimate a scale's temporal stability reliability. Using more modern scoring approaches offers a more acceptable method, but even in this instance, assigning different scale values to items based on earlier phases of scale construction is risky. (The interpretation or perceived extremity of items may change across measurement periods.)

In addition to the fact that this scaling model forces the investigator to employ the more costly test-retest technique to estimate reliability, there are other methodological issues that should be considered when using this approach. For example, whether a judge can be sufficiently objective to disregard important personal feelings in evaluating the favorability of an attitude item is an open question. However, Webb (1955) and Upshaw (1965) argued that although the absolute score assigned an item may vary as a function of judges' attitudes, and from time to time, the rank order of the items is maintained – i.e., the *relative* position of items remains unchanged no matter what the judges' attitudes – and thus the utility of Thurstone's stimulus scaling procedure in developing scales of individual differences is maintained. Although this issue is unresolved, there is little disagreement about the fact that constructing Thurstone scales is difficult and time-consuming, and, further, that such scales do not take advantages of available technological developments. In the sections that follow, we will discuss some of the more popular of the alternative models used in scaling differences among individuals.

Guttman's Scalogram

An early attempt to improve upon the Thurstone model was suggested by Louis Guttman (1944, 1947; Guttman & Suchman, 1947). The Guttman scalogram method makes use of the concept of *cumulative* or *monotone* items. With items of this type, the more favorable (or extreme) the respondent's attitude toward an object or issue, the higher (or more extreme) the individual's total attitude score. Nunnally's (1967) earlier example of tests of spelling ability here is informative.

TABLE 16.2 Example of Guttman Scalogram for “Scale of Attitudes toward Public Health Care”

Item #	Item	Yes	No
1	A universal public health care plan might in the long run prove to be beneficial in America.		
2	It is probably a good idea that the United States begin a universal public health care plan.		
3	A universal public health care plan is in the best interests of the country.		
4	A universal public health care plan would be a very positive development in the United States.		
5	A universal public health care plan would be the best thing that ever happened to people in the United States.		

Presumably, a child who could spell a difficult word would have little trouble with less difficult ones. Similarly, a child who had trouble spelling even easy words would find spelling difficult ones next to impossible. The idea of a cumulative or monotonically increasing level of difficulty (or extremity of belief) underlies this approach. The hallmark of Guttman’s scalogram method is that it presents participants with items of increasing extremity on the issue under investigation. If the scale is of high quality, an individual who endorses an item at a given level of extremity (or favorability) also is expected to endorse all less extreme items. Under ideal conditions, knowledge of a participant’s total score would enable the investigator to reproduce exactly the individual’s pattern of responses. Consider the hypothetical scale of Table 16.2.

Items that make up the scale are arranged in a gradually ascending order, with item 5 being the most positive evaluation of (or attitude toward) public health care in the United States. If the scale is found to be reliable (or reproducible, to use Guttman’s term), we would expect that those respondents who endorsed item 3 would also have endorsed items 1 and 2. Further, the knowledge that a respondent endorsed two items should enable us to predict with a high degree of certainty that the chosen items were items 1 and 2 (that is, if the scale was highly reproducible). If an investigator can reconstruct the specific items a respondent checked by knowing the total number of items endorsed, the scale is said to possess a high **coefficient of reproducibility**. To determine the coefficient of reproducibility, the statistical expression of the extent to which participants’ patterns of response can be inferred from their total scores, we need to know the total number of responses generated by the total sample of respondents, and the number of times participants’ choices fell outside of the predicted pattern of responses (errors). To calculate this statistic, use the following formula:

$$\text{Coefficient of Reproducibility} = 1 - (\text{Total Errors} / \text{Total Responses})$$

The scale construction procedures used in Guttman’s scalogram are all designed to result in a scale with a high coefficient of reproducibility. It often is erroneously assumed that a high coefficient value indicates a unidimensional scale. However, if the likelihood of endorsement (i.e., the “popularity” of an item) varies greatly from item to item, it is possible to obtain a high coefficient of reproducibility with items that have very little to do with one another. Imagine that the scale consisted of items on public health care policies as well as liberal political ideology. Generally, people who support universal public health care tend to hold more politically liberal views. This combination would result in a high coefficient of reproducibility, even though the items tap separate constructs.

Guttman's scalogram approach is underutilized because of the difficulty involved in establishing a trustworthy (internally consistent) reliability coefficient. Green (1956), Cliff (1977), and Kenny and Rubin (1977) have all discussed this issue and have proposed alternate and generally more conservative methods of assessing the reproducibility of Guttman scales. Although these alternatives are improvements over the standard method, none yet has gained widespread acceptance. As such, many attitude researchers avoid Guttman's approach unless it is clearly suggested by the research operations.

An exception to this general observation is found in the social distance measure devised by Bogardus (1933, 1992; see also Campbell, 1952). This measure is still used to assess the extent of social distance people would be most comfortable maintaining between themselves and a member of some identified group (Catholic, Armenian, dockworker, etc.). In this scale, respondents are asked if they would accept a member of the identified group as visitors to their country, as citizens, neighbors, close personal friends, or close relatives by marriage. Presumably, those willing to accept a person from a specific social grouping as a close friend would also be willing to allow all the lower levels of interpersonal relationships – as a neighbor, a citizen, and as a visitor to the country. Most recently, the approach has been used to assess social distancing in a pandemic (An et al., 2021; Prachthauser et al., 2020) and changes in sentiment toward various ethnic and religious groups in the United States (Boyd, 2021; Parrillo & Donoghue, 2005).

Likert's Method of Summated Ratings

The model of scale construction designed by Rensis Likert (1932) is one of the two most popular approaches for generating reliable scales to examine differences among people (Jebb, 2021). Compared to Thurstone's method of equal-appearing intervals, or Guttman's scalogram analysis, Likert's model proves not only more efficient in terms of time and resource expenditure, but also more effective in developing scales of high reliability (in terms of both internal consistency and temporal stability). Unlike Thurstone's or Guttman's methods, in which participants typically offer only a binary response (e.g., yes or no) for each item, Likert scale respondents indicate the extent of agreement or disagreement to each item using a multiple-choice format. On each item, respondents pick one of (usually) five options indicating the extent to which they agree with the position espoused in the item. Response options commonly presented are "strongly agree," "agree," "neutral," "disagree," and "strongly disagree."

In a scale of attitudes toward the Army, for example, participants might be asked to respond to the statement, "The U. S. military has been a positive force for peace throughout the world," through choosing one of these five response options. Presumably, a person with a favorable attitude toward the military would "agree" or "strongly agree" with the statement, whereas an individual with a negative view of the Army would be more likely to endorse the "disagree" or "strongly disagree" option. If we assign values of one to five to these response options (with higher scores representing more positive attitudes), then the overall attitude toward a given issue or entity would be represented by the sum or mean of the individual's responses over all the items on the scale.

The item employed here is an example of a *positively worded* statement, because choosing "agree" or "strongly agree" indicates a favorable attitude toward the object – in this instance, the military. An unfavorably, or *negatively worded* item is one on which strong agreement indicates a strong negative attitude (and with these items, the scoring procedure is reversed, i.e., "strongly agree" is scored +1, whereas "strongly disagree" is given the score of +5). An example of an item that reverses the intent of the previous example might be, "The Army has had a harmful effect on world peace."

The summation or averaging process across items is used to calculate a composite scale score, which is an implicit recognition that any single item is at best a fallible representative of the underlying construct or attitude it is intended to represent. By combining a participant's responses over many such items, however, we hope to minimize the "noise" or measurement error that the imperfections of each item contribute to the overall score (especially if the items have different sources of measurement error), thereby arriving at a more internally consistent measure of the construct.

If we consider each item in a Likert scale as a different operational definition of the attitude or construct it is intended to tap, then the logic of this scaling approach is consistent with the logic of multiple operationism. We assume that all of the "operations" (i.e., items) will miss the mark to some extent (that is, no single item will perfectly capture the construct it is intended to represent), but we attempt to design items so that they miss the mark in different ways; thus, the resulting scale (i.e., the total or averaged score across all items) should provide a more certain identification of the construct of interest than any single item. If each item is susceptible to error, and if the error is random, then over a series of items, the error should cancel, leaving only the reliable true score. The scale construction process developed by Likert is adopted to minimize effects of item-specific "irrelevancies" (error), and thereby arrive at the best operationalization of people's evaluations on any given issue (see Crano, 2000).

The initial steps of scale development using this method resemble those of both Thurstone and Guttman. As in these earlier approaches, many potential items are initially gathered, and obviously double-barreled, ambiguous, or confusing items are rewritten or discarded. At this point the similarity to earlier methods ends, for rather than searching for items that represent the entire continuum of the construct, the Likert model calls for creating items that are *moderately* favorable or unfavorable toward the attitude object under study. Because respondents indicate their *degree* of agreement using the response options for each item, generating items of widely varying degrees of favorability is unnecessary. The response format itself provides the respondent's indication of extremity. Given the scale construction procedures employed in developing Likert-type scales, the use of extreme items in the initial research phases would be a waste of effort, because the later scaling operations would almost certainly indicate that such items should be discarded. To examine the topic of parental attitudes toward spanking children as a form of discipline, for example, an appropriate item for the Likert technique would be "Parents should consider spanking as a form of discipline," using options from 1 (strongly disagree) to 5 (strongly agree). Extreme items such as "Parents have the right to beat their children" does not lend itself to Likert's scaling approach.

After creating many items that appear to tap the construct of interest, the researcher administers the item set to a large group of respondents (at least ten respondents per item, with a minimum of 200 for purposes of factor analysis). After collecting participants' responses, the researcher may use either classical measurement theory (e.g., item-total correlations) or more modern measurement approaches (e.g., confirmatory factor analysis and item response theory) to evaluate the psychometric properties of the scale (see Chapter 3 for a review of these approaches).

Frequently, item scores are summed over each participant, thereby creating a total score for each person, and then the complete matrix of intercorrelations between all pairs of items and between each item and the total score is calculated. Statistical software is used to compute the correlation matrix, which provides auxiliary information to allow for the calculation of coefficient alpha as an index of internal consistency (Chapter 3). Higher correlations generally yield a higher Cronbach's alpha. However, it is important to realize that Cronbach's alpha produces a higher coefficient as the correlation across items becomes stronger *and* as the number of items

increases. Thus, using many items in this initial scale construction phase will result in a strong alpha coefficient if the choice of items was reasonable.

The investigator's primary job at this point is to retain the items that form the best scale, and discard items that correlate poorly with the rest of the scale. Coefficient alpha, an estimate of the internal consistency of the entire set of items, is not useful in an item-by-item analysis of this type. Of more practical utility is the investigation of each item's correlation with the total score, known as the item-total correlation. As noted, the total score is conceptualized as the best estimate of the construct under investigation. However, because the total score is the outcome of many items, some of which probably are of low quality, it is far from a perfect representation of the underlying construct. To improve the precision of the scale, the investigator may discard the least fitting items; that is, those that do not correlate strongly with the total score. In developing the final scale, the researcher retains items having the highest item-total score correlations.

After having decided on the best items, and discarded the worst ones, it is necessary to recalculate the item-total correlation using the "reduced set" of items, because the total composite score changes every time an item is discarded. If the initially strong correlations are maintained, or improved, the investigator then should recalculate coefficient alpha on the reduced set to determine the degree of interrelatedness of items. An alpha coefficient of .70 or higher suggests that the scale is reasonably internally consistent. However, some "shrinkage" in the reliability coefficient must be expected when the item set is re-administered to another group, because this scale construction process capitalizes on sample-specific variations (i.e., error). The extent of such attenuation or shrinkage is usually not severe unless the new sample is very different from that on which the scale was developed originally.

If coefficient alpha is weak (for example, if it falls short of an arbitrary value of .70), the internal consistency can be improved by the addition of more items that correlate positively with the original set and with the total score. This item adding process can be continued until the desired level of reliability has been reached. It sometimes happens that a researcher has developed a multidimensional scale, which taps more than a single factor or subscale. This overall scale should consist of subscales that are somewhat correlated but are conceptually distinct. When this is anticipated to occur, a factor analysis can be employed to illuminate the underlying factors representing the multidimensional scale.

Factor analysis as an aid to scale construction. Factor analytic techniques may be divided into two types. **Exploratory factor analysis (EFA)** is a statistical technique to assess the multidimensional structure of a scale if the researcher does not have hypotheses regarding the number and types of underlying factors that will emerge in the solution. Application of EFA identifies the subset of items that are most strongly correlated, with such items forming a factor, and can allow for (and identify) multiple factors.

Confirmatory factor analysis is a statistical technique used to assess the multidimensional structure of a scale if the researcher has hypotheses regarding the number and types of underlying factors expected to emerge in the solution. A hypothesis-based approach to factor analysis requires reviewing the relevant prior literature to make informed decisions about how to phrase the items to represent each of the dimensions. In applications of confirmatory factor analysis, the investigator must force or specify in advance the proposed number of factors and test whether they fit the respondents' data. Only items that are hypothesized (without peeking at the collected data) to represent a factor should be allowed to load on that factor. Once the optimal items are retained, if the desire is to cross-validate the scale, it is then administered to a new set of participants. (See Chapter 3 for further description of these two forms of factor analysis.)

Osgood's Semantic Differential

Although Likert's approach represented an important technical advance over both the Thurstone and Guttman methods, it shares some of the liabilities of these procedures. All three scaling models require considerable expenditures of time and effort in the scale construction process, and all three techniques require the development of a new set of rating items each time attitudes toward a new person or object are to be assessed. For these reasons, a technique pioneered by Osgood and his colleagues (Osgood et al., 1957; Osgood, 1962; Snider & Osgood, 1969) has become popular as a standardized form of assessing attitudes toward objects and issues. The original development of this scaling model was stimulated by Osgood's attempts to determine the subjective meanings people attach to words or concepts. In Osgood's semantic differential scale, rather than asking respondents to respond to variations of statements concerning the concept under study (as in the Likert or Thurstone approaches, for example), the concept is presented directly and participants are instructed to react to it in the form of ratings on several bipolar adjectives (these are antonyms, or logical opposites).

In the semantic differential measurement method, a single stem concept (a word or phrase, such as *The U.S. Congress*) indicating the object to be judged is presented, followed by a set of adjective pairs to be used in the judging process. This contrasts with Likert scaling, where different stems are used to describe the item to be evaluated, but the same response options (strongly agree, agree, etc.) are consistently used to capture responses. To illustrate the distinction, suppose we constructed two scales to measure "Attitudes toward my mother" using the Likert and Osgood models. These scale types are contrasted in Table 16.3. They each make use of five items, with each item anchored on 5-point options. With the Likert scale, it is necessary to generate a diverse list of statement stems, using the same "strongly agree" to "strongly disagree" response options throughout. In contrast, using the semantic differential model requires presenting the construct, "My Mother," followed by different bipolar adjectives (which represent different items in Osgood's measurement model). Responses closer toward the middle of each set of bipolar adjectives reflect more neutral evaluations on the continuum, whereas those at the ends of the scale represent more extreme evaluations, either positive or negative. Examples of both rating approaches are presented in Table 16.3.

A different sentence stem for each item is not required in Osgood's approach. Another advantage is that a variety of different constructs could be used later to create other semantic differential evaluations, using the same "items." That is, the same set of bipolar adjectives shown in Table 16.3 could be used subsequently to construct measures of attitudes toward father, best friend, or even romantic partner. Another obvious advantage concerns time savings for participants and scale constructors. Even with the same number of items, a semantic differential scale will be shorter in length, easier to read, more difficult to misinterpret, and therefore completed more quickly by respondents than a Likert scale. From the scale constructor's perspective, the Osgood approach is considerably more efficient.

Many of the statistical analyses used to evaluate the properties of Likert scales are used in evaluating semantic differential scales. Ratings made on any of the adjective anchors could be correlated with those made on each of the other adjective anchors (or semantic differential items, to use another term), and these data may be used in statistical analyses such as item-total correlations and coefficient alpha. The use of factor analysis will provide information regarding the extent to which various bipolar adjectives are clustered and are independent from other items. Items that "load on the same factor" are highly interrelated with one another, and relatively weakly related with other items that do not load on their factor. It is assumed that items that cluster together are focused on the same underlying psychological dimension or construct.

TABLE 16.3 Construction of scales of “attitudes toward mother” with Likert and Osgood models

<i>Method of summated ratings (Likert)</i>					
1.	I think my mother is a good person				
	Strongly disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly agree (5)
2.	My mother is a kind person				
	Strongly disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly agree (5)
3.	I believe my mother is generous				
	Strongly disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly agree (5)
4.	My mother has a pleasant demeanor				
	Strongly disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly agree (5)
5.	I feel that my mother tries to be fair with me				
	Strongly disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly agree (5)
<i>Semantic Differential Scale (Osgood)</i>					
	My Mother				
1.	Bad (1)	(2)	(3)	(4)	Good (5)
2.	Cruel (1)	(2)	(3)	(4)	Kind (5)
3.	Stingy (1)	(2)	(3)	(4)	Generous (5)
4.	Unpleasant (1)	(2)	(3)	(4)	Pleasant (5)
5.	Unfair (1)	(2)	(3)	(4)	Fair (5)

Several studies have examined the dimensions of common adjective anchors (see Osgood & Luria, 1954; Osgood et al., 1957; Snider & Osgood, 1969) with widely varying concepts being employed, and with respondents from 26 different cultures around the world. Consistent findings have emerged: Three factors or clusters of adjectives have emerged regularly in factor analytic studies of the semantic differential: (1) *evaluation* or value of the object (e.g., good/bad), (2) *potency* or power of the object (e.g., strong/weak), (3) and *activity* or dynamism of the object (e.g., slow/fast). The adjectives involving evaluations tend to account for the most meaning, or the largest proportion of the factor variance in respondents' subjective interpretations of objects. Although it is most ideal to develop a semantic differential scale with items incorporating bipolar adjectives involving a mix of evaluation, potency, and activity, the exclusive use of evaluative anchors is common, and sufficient for most research purposes. Thus, for example, to use semantic differential scales to measure people's attitudes toward public health care, we might employ the evaluative adjective pairs of Figure 16.1. The overall attitude in this case is defined as the sum or average of a respondent's scores over all ten of the semantic differential evaluative items when rating the concept (in this case, public health care).

In this measurement approach, a respondent checks the point on each item that best indicates the degree of positivity or negativity toward the concept in question. It is common that other types of response anchors are interspersed among the critical evaluative items. Items with response formats connoting potency (e.g., strong-weak, rugged-delicate, large-small, hard-soft, heavy-light, etc.), and activity (e.g., active-passive, quick-slow, sharp-dull, excitable-calm, hot-cold) could be included, for example. However, in defining respondents' attitudes, only the responses on the evaluative items are summed. The summation process is identical to that involved in Likert scaling – and it involves the same assumptions, dangers, etc. Usually, 7-point item response formats are used in semantic differential research, with higher scores used to connote positive views toward the object.

The semantic differential approach offers many practical advantages. Given the nature of the statistical process through which the various factors or “clusters” of items were developed, it is

Item #	Public Health Care									
1	Good	:	_____	:	_____	:	_____	:	_____	: Bad
2	Kind	:	_____	:	_____	:	_____	:	_____	: Cruel
3	Beautiful	:	_____	:	_____	:	_____	:	_____	: Ugly
4	Pleasant	:	_____	:	_____	:	_____	:	_____	: Unpleasant
5	Fair	:	_____	:	_____	:	_____	:	_____	: Unfair
6	Honest	:	_____	:	_____	:	_____	:	_____	: Dishonest
7	Clean	:	_____	:	_____	:	_____	:	_____	: Dirty
8	Valuable	:	_____	:	_____	:	_____	:	_____	: Worthless
9	Positive	:	_____	:	_____	:	_____	:	_____	: Negative
10	Wise	:	_____	:	_____	:	_____	:	_____	: Foolish

FIGURE 16.1 Semantic differential items measuring “attitudes toward public health care.”

safe to assume that the internal consistency of such a measurement instrument will be high, an assumption that always should be verified. Investigations of the temporal stability of such instruments also have provided strong support for this measurement approach (e.g., Osgood et al., 1957; Snider & Osgood, 1969). The generality of the evaluative response as a major component of an individual’s subjective reaction toward any object also is advantageous. The semantic differential technique apparently offers the researcher a ready-made attitude scale for assessing the beliefs and attitudes of almost any topic (see Osgood & Luria, 1954, 1976). As such, it offers a great practical advantage over the other forms of attitude assessment, all of which demand greater time expenditures in instrument development. On the downside, the semantic differential provides information only on affect toward the object (stem), with no additional content. The continued popularity of the semantic differential as a measurement instrument for attitudes and other constructs is evident in the literature.

Response Format

Care and common sense should be exercised when choosing the specific response adjectives for items used in a semantic differential instrument, because the specific attitude object under investigation could affect the meaning or appropriateness of the scales employed. For example, the bipolar adjectives of fair–unfair appear appropriate if used to determine people’s attitudes toward a Justice of the U.S. Supreme Court; these same response anchors, however, would be less than optimal if the object of judgment were Ben & Jerry’s Cherry Garcia Ice Cream. In this case, the use of an inappropriate rating introduces unnecessary error or imprecision in capturing participants’ evaluations. To detect the presence of this problem (because, sometimes, inappropriate item stem and response anchor pairings are difficult to recognize), it is wise to calculate coefficient alpha and item–total correlations on all scales using semantic differential items. A strong coefficient of internal consistency and high item–total correlations indicate the items chosen to measure the attitude object are related to the overall composite.

In the Thurston and Guttman scale construction techniques, each item requires participants to select from a binary option (e.g., “yes or no,” “true or false”) to indicate their endorsement of an item. The Likert and Osgood techniques use more continuous response formats (e.g., “strongly agree” to “strongly disagree”). Technically, Likert and Osgood scale items use

an ordinal response format; that is, the psychological distance between “strongly disagree” to “disagree” is not necessarily the same as the distance between “disagree” to “neutral,” or “neutral” to “agree.” These response options have the feature of rank order, but not equal unit difference between response options. To satisfy the criterion for interval scaling, equal distances between any two adjacent response categories are required, but it cannot be presumed in Likert or Osgood-type scales. Fortunately, research suggests it is acceptable to apply many parametric statistical techniques (based on the assumption of a normal distribution) to the outcomes of Likert and Osgood scaling procedures (Norman, 2010). Statistical simulation research indicates that ordinal responses, especially if overall composites are used in the statistical analyses, approximate and satisfy distributional properties of a normal distribution.

Number of response options. Another response format issue requiring consideration concerns the number of response categories appropriate for an item. Items consisting of five and seven response categories are the most encountered, but some researchers use a greater number, such as 10-point rating scales (Wakita et al., 2012). Incorporating a greater number of response options can result in more gradations in responding, which may result in a more continuous assessment of a construct. Increasing the number of response categories thus may be beneficial in making the item more sensitive in discerning and differentiating smaller distinctions among participants. A trade-off in the inclusion of a greater number of response options is the greater amount of time required to contemplate which of the many categories best corresponds to one’s attitude toward an issue, and there is some question regarding respondents’ capacities to make fine-grained distinctions in responding to attitude items. Although raising the number of response categories is beneficial, these advantages are relatively minor (Wakita et al., 2012). Malhotra et al. (2009) have suggested that 7-point ratings offer the optimal level of discriminability.

Bipolar versus unipolar response ratings. Another factor that can influence the number of response options appropriate for a particular rating is whether the rating is *bipolar* (going from negative at one anchor to positive at the other) or not. Most Likert-type scales use a bipolar response format (strongly disagree to strongly agree), and semantic differential ratings also tend to be bipolar (e.g., bad–good). However, some rating dimensions are more likely to be *unipolar*, ranging from low (e.g., not at all) to high (e.g., very much). Response scales that involve estimates of frequency would generally be unipolar. Some types of ratings could be either unipolar or bipolar. For example, if an item asked respondents to rate how “just” or “fair” a particular policy is, the rating could be made on a scale from “very unjust” to “very just” or just in terms of degree of justness: “not at all just” to “completely just.” As a rule, unipolar ratings should be limited to 5-point options whereas bipolar ratings can be either 5-point or 7-point rating scales.

Anchor extremity. Another factor affecting choice of response format is the extremity of endpoint labels (Wyatt & Meyers, 1987). Research indicates that if the labels of the anchor points are less extreme (e.g., 1 = “mostly false” to 5 = “mostly true”) rather than more extreme (1 = “completely false” to 5 = “completely true”), participants tended to select each of the possible options more equally. When using extreme adjectives for anchor labels, endpoints were rarely selected, and most responses were distributed toward the middle, resulting in less variability in responses. Given that a practical scale should be effective in capturing a wide range of responses about an object, it is sound advice to avoid the use of extreme endpoint labels in most scales. However, it might be practical to use extreme labels in research domains on which people tend to have strong polarizing beliefs.

Use of Reversed Items

Another consideration concerns the choice of using or excluding reversed items. Many scales contain a mixture of positively and negatively phrased (or reversed) items. Reversed items are

used to guard against the threat to measurement validity known as *acquiescence response bias*, the penchant for some individuals to agree with positively worded item statements regardless of content. Acquiescent responding could occur because of a personality disposition to be agreeable or the desire to please the investigator. The positively worded item of “Creative thinking promotes good decisions in my daily life” may be rephrased in the form of a negation of the positive statement, “Creative thinking does not promote good decisions in my daily life,” or in the form of an opposing statement, “Creative thinking interferes with good decision making in my life.” In computing the overall score, the reversed items should be scored in the opposite direction from the positively worded items when using a response format of “strongly agree” to “strongly disagree.” At face value, both statements appear to be nothing more than simple opposites, but item reversals produce greater misinterpretation, because such statements are more cognitively challenging to process and judge (Swain et al., 2008; Weijters et al., 2010). Negatively worded item stems require greater capacity in working memory, and greater focus to interpret, leaving less room for thought. Misunderstanding reversed items arises due to additional complexity in interpreting the meaning of the sentence when including the word “not.” Heuristically, people tend to store beliefs in memory in the affirmative direction, not in terms of its negated counterpart. Furthermore, especially in lengthy scales, some participants may fail to attend to every single word in every statement, accidentally skipping over encounters with “not.”

Investigators’ desires to minimize participant acquiescence should be weighed against participant challenges in interpreting items correctly. Some research has suggested that scales should only use affirmative item stems, as the problems of interpretation introduced by reversals outweigh the problems attributed to acquiescence (Ebesutani, 2013). Although no definitive solution exists to simultaneously protect against acquiescence and cognitive load, a possibility is to keep the phrasing of all item stems in the affirmative direction but reverse the direction of response options for some items (Barnette, 2000). For example, create half the items with response options listed starting from the left with “strongly disagree” to “strongly agree,” with the remaining items using reversed options listed from “strongly agree” to “strongly disagree.” Afterward, all responses should be coded in the same direction, with higher values representing greater degree of that construct, so that a mean composite may be computed subsequently. Research has demonstrated that scales with a mix of items with reversals and non-reversals of option anchors yield similar Cronbach’s alphas to scales containing all item stems and responses phrased in the same direction, and both strategies generate better reliability coefficients than scales containing a mixture of items with reversed and non-reversed stems (Barnette, 2000). In their comparison of semantic differential and Likert-based measures, Friborg et al. (2006) found support for Osgood’s approach and suggested that its format avoided the acquiescence problem. Both observations argue in favor of the semantic differential scaling approach.

Conclusion

The systematic measurement of people’s underlying beliefs, knowledge, attitudes, and values represents a major achievement of social science research. Owing to the imagination of our psychometric forebears, and the technical virtuosity of those that followed, we have developed in the social sciences a set of techniques that allow us to peer into the thought processes of cooperative (and sometimes uncooperative) respondents. Scientifically grounded rating scale methodologies represent one of our field’s greatest achievements. Properly applied, well-designed questionnaires and scales allow us some access to the thoughts and views of our respondents and help us understand the views of the person or group under observation. Questionnaires and scales are powerful tools for applied and basic research. From scientific polling techniques to

theory-testing research, researchers rely heavily on these self-report instruments as a versatile component of our methodological tool chest.

Notes

- 1 Some researchers even recommend that the order of response options be rotated (e.g., see Krosnick, 1991; Sudman et al., 1996). This recommendation is made because of some evidence that suggests that people tend to ascribe more to the initial response options on written questionnaires, and to later options on questionnaires that are read to them by an interviewer (Krosnick & Alwin, 1987; McClen-don, 1991).
- 2 The equality of intervals separating the scale values of items is an important criterion of Thurstone's technique. If this requirement is met, it suggests that the scale may be treated as being of interval, rather than ordinal level, and this has implications for the most appropriate type of statistical analysis that may be used.
- 3 Note that in these scales, higher parenthesized values represent less positive attitudes. Thus, an item score of 11 means that all judges categorized the scale as the most negative score that could be provided.

References

- An, L., Hawley, S., Van Horn, M. L., Bacon, E., Yang, P., & Resnicow, K. (2021). Development of a coronavirus social distance attitudes scale. *Patient Education and Counseling*, 104(6), 1451–1459.
- Barnette, J. J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using these negatively worded stems. *Educational and Psychological Measurement*, 60, 361–370.
- Beatty, P., Herrmann, D., Puskar, C., & Kerwin, J. (1998). 'Don't know' responses in surveys: Is what I know what you want to know and do I want you to know it? *Memory*, 6(4), 407–426.
- Bogardus, E. S. (1933). Social distance and its practical applications. *Sociology and Social Research*, 22, 462–476.
- Bogardus, E. S. (1992). Social distance and its origins. *Journal of Applied Sociology*, 9, 216–226.
- Boyd, R. L. (2021). Social distance reexamined: European ancestry groups in the early twentieth-century United States. *Sociological Inquiry*, 91(3), 559–577.
- Campbell, D. T. (1952). The Bogardus social distance scale. *Sociology & Social Research*, 36, 322–326.
- Cliff, N. (1977). A theory of consistency of ordering generalizable to tailored testing. *Psychometrika*, 42, 375–399.
- Converse, P. E. (1964). The nature of belief systems in mass publics. In D. Apter (Ed.), *Ideology and discontent* (pp. 206–261). Glencoe, IL: Free Press of Glencoe.
- Coombs, C. H. (1950). Psychological scaling without a unit of measurement. *Psychological Review*, 57, 145–158.
- Crano, W. D. (2000). The multitrait-multimethod matrix as synopsis and recapitulation of Campbell's views on the proper conduct of social inquiry. In L. Bickman (Ed.), *Research design: Donald Campbell's legacy* (Chapter 3, pp. 37–61). Thousand Oaks, CA: Sage.
- Crano, W. D., & Lac, A. (2012). The evolution of research methodologies in (social) psychology. In A. Kruglanski & W. Stroebe (Eds.), *Handbook of the history of social psychology* (pp. 159–174). New York, NY: Psychology Press.
- Ebesutani, C. K. (2013). Evidence based assessment strategies for 'real world' clinical settings. 73, ProQuest Information & Learning, US. Retrieved from <http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2013-99060-306&site=ehost-live&scope=site>
- Fabrigar, L. R., Krosnick, J. A., & MacDougall, B. L. (2005). Attitude measurement: Techniques for measuring the unobservable. In T. C. Brock & M. C. Green (Eds.), *Persuasion: Psychological insights and perspectives* (2nd ed., pp. 17–40). Thousand Oaks, CA: Sage.
- Friborg, O., Martinussen, M., & Rosenvinge, J. H. (2006). Likert-based vs. semantic differential-based scorings of positive psychological constructs: A psychometric comparison of two versions of a scale measuring resilience. *Personality and Individual Differences*, 40(5), 873–884.
- Green, B. F., Jr. (1956). A method of scaleogram analysis using summary statistics. *Psychometrika*, 21, 79–88.

- Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, 9, 139–150.
- Guttman, L. (1947). The Cornell technique for scale and intensity analysis. *Educational and Psychological Measurement*, 7, 247–279.
- Guttman, L., & Suchman, E. A. (1947). Intensity and zero point for attitude analysis. *American Sociological Review*, 12, 57–67.
- Kenny, D. A., & Rubin, D. C. (1977). Estimating chance reproducibility in Guttman scaling. *Social Science Research*, 6, 188–196.
- Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5, 213–236.
- Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response order effects in survey measurement. *Public Opinion Quarterly*, 51, 201–219.
- Krosnick, J. A., & Fabrigar, L. R. (2001). *Designing great questionnaires: Insights from psychology*. Oxford, UK: Oxford University Press.
- Krosnick, J. A., Holbrook, A. L., Berent, M. K., Carson, R. T., Hanemann, W. M., Kopp, R. J., ... & Conaway, M. (2002). The impact of ‘no opinion’ response options on data quality: Non-attitude reduction or an invitation to satiate? *Public Opinion Quarterly*, 66, 371–403.
- LeCount, R. J. (2022). Who wants to know? White racial attitudes and survey item refusal on gun ownership questions. *Sociological Inquiry*, 92(1), 153–199.
- Lee, S.-G., & Kanazawa, Y. (2000). Handling ‘don’t know’ survey responses: The case of Japanese voters on party support. *Behaviormetrika*, 27, 181–200.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22 140, 55.
- Luskin, R. C., & Bullock, J. G. (2011). ‘Don’t know’ means ‘don’t know’: DK responses and the public’s level of political knowledge. *The Journal of Politics*, 73, 547–557.
- Magelssen, M., Supphellen, M., Nortvedt, P., & Materstvedt, L. J. (2016). Attitudes towards assisted dying are influenced by question wording and order: A survey experiment. *BMC Medical Ethics*, 17, 1–9.
- Malhotra, N., Krosnick, J. A., & Thomas, R. K. (2009). Optimal design of branching questions to measure bipolar constructs. *Public Opinion Quarterly*, 73(2), 304–324.
- McClendon, M. J. (1991). Acquiescence and recency response-order effects in interview surveys. *Sociological Methods and Research*, 20, 60–103.
- Norman, G. (2010). Likert scales, levels of measurement and the “laws” of statistics. *Advances in Health Science Education*, 15, 625–632.
- Nunnally, J. C. (1967). *Psychometric theory*. New York: McGraw-Hill.
- Osgood, C. E. (1962). Studies on the generality of affective meaning systems. *American Psychologist*, 17, 10–28.
- Osgood, C. E., & Luria, Z. (1954). Applied analysis of a case of multiple personality using the semantic differential. *Journal of Abnormal and Social Psychology*, 49, 579–591.
- Osgood, C. E., & Luria, Z. (1976). A postscript to “The three faces of eve.” *Journal of Abnormal Psychology*, 85, 276–286.
- Osgood, C. E., Suci, D. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois Press.
- Parrillo, V. N., & Donoghue, C. (2005). Updating the Bogardus social distance studies: A new national survey. *The Social Science Journal*, 42(2), 257–271.
- Prachthauser, M., Cassisi, J. E., Le, T.-A., & Nicasio, A. V. (2020). The social distance scale (v1): A screening instrument to assess patient adherence to prevention strategies during pandemics. *International Journal of Environmental Research and Public Health*, 17(21), 8158; Nov 4. <https://doi.org/10.3390/ijerph17218158>
- Rasinski, K. A., Lee, L., & Krishnamurty, P. (2012). Question order effects. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology: Foundations, planning, measures, and psychometrics* (Vol. 1, pp. 229–248). Washington, DC: American Psychological Association.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.
- Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Cambridge, MA: Academic Press.
- Snider, J. G., & Osgood, C. E. (Eds.) (1969). *Semantic differential technique*. Chicago, IL: Aldine.
- Standing, L. G., & Shearson, C. G. (2010). Does the order of questionnaire items change subjects’ responses? An example involving a cheating survey. *North American Journal of Psychology*, 12, 603–614.

- Stouffer, S. A. (Ed.) (1950). *Measurement and prediction*. Princeton, NJ: Princeton University Press.
- Sudman, S., & Bradburn, N. M. (1982). *Asking questions: A practical guide to questionnaire design*. San Francisco, CA: Jossey-Bass.
- Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco, CA: Jossey-Bass.
- Swain, S. D., Weathers, D., & Neidrich, R. W. (2008). Assessing three sources of misresponse to reversed Likert items. *Journal of Marketing Research*, 45, 116–131.
- Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, 33, 529–554.
- Thurstone, L. L. (1931). The measurement of attitudes. *Journal of Abnormal and Social Psychology*, 26, 249–269.
- Thurstone, L. L., & Chave, E. L. (1929). *The measurement of attitudes*. Chicago, IL: University of Chicago Press.
- Tourangeau, R., Couper, M. P., & Conrad, F. (2004). Spacing, position, and order: Interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly*, 68, 368–393.
- Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, 103, 299–314.
- Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. New York, NY: Cambridge University Press.
- Upshaw, H. S. (1965). The effects of variable perspectives on judgments of opinion statements for Thurstone scales: Equal appearing intervals. *Journal of Personality and Social Psychology*, 2, 60–69.
- Visser, P. S., Krosnick, J. A., & Lavrakas, P. J. (2000). Survey research. In H. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 223–252). Cambridge: Cambridge University Press.
- Wakita, T., Ueshima, N., & Noguchi, H. (2012)). Psychological distance between categories in the Likert scale: Comparing different number of options. *Educational and Psychological Measurement*, 72, 533–546.
- Webb, S. C. (1955). Scaling of attitudes by the method of equal-appearing intervals. *Journal of Social Psychology*, 42, 215–239.
- Wegener, D. T., Downing, J., Krosnick, J. A., & Petty, R. E. (1995). Measures and manipulations of strength-related properties of attitudes: Current practice and future directions. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 455–487). Mahwah, NJ: Erlbaum.
- Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: the number of response categories and response category labels. *International Journal of Research in Marketing*, 27, 236–247.
- Wyatt, R. C., & Meyers, L. S. (1987). Psychometric properties of four 5-point Likert type response scales. *Educational and Psychological Measurement*, 47, 27–35.
- Yeung, J. T., Livingston, M., Callinan, S., Wright, C., Kuntsche, E., Room, R., & Dietze, P. (2020). Effects of question type and order when measuring peak consumption of risky drinking events. *Alcohol and Alcoholism*, 55(6), 631–640.

17

SCALING STIMULI

Social Psychophysics

Developing and using scales of high utility, generality, and psychometric quality (i.e., of high reliability and validity) is a common part of the social researcher's job description. As discussed in Chapter 16, questionnaires and rating scales are constructed to measure differences among individuals. For example, a scale of attitudes toward nuclear energy, assessed with multiple items, has as its purpose the classification of individuals along a continuum, so that those with the highest scores are defined as most in favor of adopting nuclear energy, and those with the lowest scores the least favorable. Those scoring in the middle are considered intermediate on their attitudes toward this energy source. In this form of scaling, the items are used to compute a summary score for the purpose of arranging people at some point on the scale: All items are designed to tap the same construct; that is, the items are assumed to differ only in terms of measurement error. So, in our attitudes toward nuclear energy scale, we assume that the questions (items) we pose to respondents, and that are used to create our summary score, all tap the same underlying construct (in this case, attitudes toward nuclear energy). Measures of this variety are called tests of differences among individuals, or more commonly, tests of individual differences. Individual difference assessments are the most common form of measures used in contemporary social research.

Although less common, we sometimes are concerned with perceived differences among a set of stimuli (i.e., to what extent stimulus 1 differs from stimuli 2 and 3), rather than differences among a set of respondents (i.e., to what extent person 1 differs from persons 2 and 3). Developing measurement scales that tap into perceived differences among stimuli (instead of differences between individuals) is called **stimulus scaling** (or *scales of stimulus differences*). As will be shown, creating stimulus scales calls for an approach and a set of assumptions different from those used in individual differences scaling.

The psychophysicists of more than a century ago attempted to bring the measurement logic of physics into the psychological laboratory – to borrow and apply the methods of measurement used in the physical sciences to the behavioral sciences. They were interested in developing measures of features for which no obvious metric existed (e.g., beauty, taste, fear of clowns, preference for violent movies, etc.), and to do so with high degrees of reliability and validity. A central goal of their efforts was to produce measures that accurately represented the judgments of the sample on which the measure was based. It is for this reason that Dawes and Smith (1985,

p. 515) referred to stimulus scales as “group attitude scales.” The real utility of the stimulus scaling approaches is that they allow us to impose a metric on judgments for which no obvious “yardstick” exists. In the physical sciences, this issue usually does not arise – if we wish to know the weight of an object, we place it on a scale; if we wish to know the loudness of a sound, we use a decibel meter. But suppose we want to order a group of people’s judgment of ten paintings, from most to least beautiful, and to do so in a way that the distances between adjacent points on our “beauty” scale are approximately equal. Or, more prosaically, suppose we wanted to have a group of friends arrange ten different brands of pizza from most to least delicious, along a “deliciousness” continuum that represented the group judgment and provided a clear indication of the distances that separated each pizza from all the others.

With these two sets of stimuli, no obvious physical yardstick exists. We do not have rulers with which we can unambiguously and with high consensus judge beauty or deliciousness. However, such measurement scales can be developed. In building such a scale, any differences in people’s orderings, which signify individual differences of opinion that could vary from person to person, are considered irrelevant. The statistical treatment of the data, collapsed across all respondents in the entire sample, reveals not only the relative popularity of each of the stimuli, but also their relative distances from one another, on a scale having interval properties (i.e., equal distances separate contiguous units across the entire continuum). In marketing research conducted by a pizza company, such findings might be profitably used to mimic the ingredients of the most popular brand, or to identify closest competitors.

In stimulus scaling we are concerned with differences between stimuli (or items). As such, systematic variations among stimuli are considered meaningful and important – they are the focus of investigation. Variations in judgments among participants are considered the result of error. This is an important distinction between stimulus scaling and individual differences scaling, and bears reemphasis. In stimulus scaling, we assume that individual differences among respondents’ perceptions and evaluations of the stimuli being judged are the result of error. Differences in respondents’ judgments are not viewed as the result of meaningful differences among people. Recall the opposite assumption is made in scaling individuals – that is, we assume that there are no differences among the stimuli (or items) that constitute the measurement instrument. In scales of individual differences, all items are assumed to tap the same underlying construct – belief, trait, attitude, norm, intention; variations across respondents are real and potentially meaningful. These differences in assigning meaningfulness to judgments differentiate the two scaling approaches. They indicate the two scaling approaches can and should be applied to different types of research problems, and will produce different outcomes and different ways of arranging stimuli and people. Each type of scaling method is valuable in its own right, and each satisfies different research needs.

Scaling Stimuli

Although judgments obtained from human respondents are used to construct scales of stimuli and scales of individual differences, the approaches require development of different measurement operations. A typical use of stimulus scaling techniques is found in marketing research, where a researcher may be interested in comparative evaluations of different products, brands, packages, etc. Another use is found in the political arena, where comparative evaluations of competing candidates or social policies might be under scrutiny, or in criminology, where the perceived seriousness of different types of crimes might be at issue.

Typically, in stimulus scaling studies, respondents compare one stimulus against all the others along a specific dimension or quality, or judge stimuli in terms of their degree of similarity or

dissimilarity. These comparisons are used to form a continuum, choice matrix, or perceptual map that represents the aggregate of participants' judgments regarding relative similarities of the objects in the stimulus set (Hout et al., 2013).¹ Using this technique, we obtain an ordering of stimuli along the continuum, and the intervals along the continuum are equal – in other words, the procedure produces scales of equal unit interval quality.

In one of the classic stimulus scaling investigations, Louis Thurstone (1927) assessed the beliefs of 266 University of Chicago students regarding the seriousness of 19 different crimes. Among others, the crimes included arson, bootlegging, forgery, homicide, larceny, libel, perjury, rape, and smuggling.² Rather than ask the students to rate the seriousness of each crime on 10-point scales, Thurstone paired every crime with each of the others, and asked his participants simply to underline the more serious of the two for every comparison. So, on a given judgment, a participant might be asked to judge whether perjury or rape was the more serious crime. On the next judgment, the participant might judge the more serious of libel versus bootlegging. Crano and Cooper (1973) have argued that such binary judgments often are more reliable, and less demanding of respondents than those that require fine-grained ratings or discriminations. Interestingly, Thurstone's results revealed that his sample of university students (in 1927) considered crimes against persons (homicide, rape, kidnapping, assault) as being the most serious, just as students today probably would. Property offenses and victimless crimes (vagrancy, receiving stolen goods, etc.) fell into the less serious segment of the scale. Later replications of this study by Coombs (1967) and Borg (1988) produced results consistent with the early findings, although the researchers did find some deviations from the original study. For example, in Coombs' (1967) study, 369 University of Michigan students judged rape the most serious offense and homicide second most serious, whereas this ordering was reversed in Thurstone's (1927) original study. Coombs' research also disclosed that students of 1967 did not consider bootlegging (illegal manufacturing of alcohol) nearly as serious as Thurstone's participants had, whereas seduction of a minor was considered more serious in the later sample. These variations between studies probably are attributable to changing attitudes over time or differences in the interpretation of various crime labels.³

Stimulus Scaling Techniques

The Method of Pair Comparison Scaling

Of all the classic psychometric stimulus scaling techniques, two approaches – the method of pair comparison and the method of rank order – are most common in social research. In the method of **pair comparison scaling**, a set of stimuli is examined by presenting every possible pair of stimuli to each respondent who chooses whether they prefer one stimulus over the other based on a quality or dimension (e.g., beauty, taste, artistic merit) stipulated by the researcher. This is the method Thurstone (1927) used in his study of the perceived seriousness of crimes. In this approach, choices are aggregated across all participants and the analysis of the data provides a summary of the combined group's order of preferences among stimuli.

To lend some degree of concreteness to this discussion, consider the following example. Suppose that we were interested in respondents' attitudes regarding relative differences in acting ability among six popular male actors. This *stimulus set* includes Brad Pitt, Tom Cruise, Johnny Depp, Leonardo DiCaprio, Morgan Freeman, and Anthony Hopkins.

Notice that we are not concerned that one respondent thinks Hopkins is a better actor than Depp, whereas another has the opposite opinion. In stimulus scaling, we are concerned with how the stimuli (our six actors) are arranged along a continuum of acting ability by the total participant

sample. Differences of opinion among the individuals who constitute the sample are not at issue. Indeed, as noted, these participant differences are considered a form of measurement error.

There is no obvious “acting ability” yardstick that we can use to order these actors, yet the task is far from impossible. Employing the method of pair comparison scaling, we would first assemble every possible pair of actors. In this case, this process would yield 15 nonrepetitive pairs.⁴ For example, Depp would be paired with Cruise, DiCaprio, Freeman, Hopkins, and Pitt; Cruise with DiCaprio, Freeman, Hopkins, and Pitt, and so on. Identical stimuli would not be paired (i.e., Depp obviously would not be paired with Depp). Then, each of the pairings would be presented to each respondent, with the instructions, “On each card, circle the better actor of each pair of actors.” To facilitate this task and avoid problems that might occur if every participant received the same ordering of pairs, we might place each pair of names on index cards, and randomly shuffle the cards prior to giving them to participants. The process of presentation and randomization of each pair can be facilitated with a computer. The entire deck of index-card stimuli might look like those of Figure 17.1.

Suppose we administered our set of 15 pair comparison cards to 100 participants. To summarize the obtained data, we could arrange respondents’ judgments in a *choice matrix*, as in Table 17.1. In this matrix, the cell entries represent the number of respondents who chose the *column* stimulus over the *row* stimulus. So, in Table 17.1, the data indicate that Depp was chosen over Cruise by 40 of 100 respondents (conversely, Cruise was chosen by 60 of 100 participants as a better actor than Depp). Freeman was chosen by an equal number of respondents when compared with Pitt. Although as noted, no pairing entailed Depp versus Depp, it is standard to enter an equal probability of selection (e.g., .50) for the diagonal values of the matrix. The column mean gives a reasonable indication of the aggregate opinion regarding the acting ability of the sample who judged the stimulus set. The mean data suggest that the respondents viewed

Johnny Depp	Tom Cruise	Tom Cruise	Leonardo DiCaprio	Leonardo DiCaprio	Anthony Hopkins
Johnny Depp	Leonardo DiCaprio	Tom Cruise	Morgan Freeman	Leonardo DiCaprio	Brad Pitt
Johnny Depp	Morgan Freeman	Tom Cruise	Anthony Hopkins	Morgan Freeman	Anthony Hopkins
Johnny Depp	Anthony Hopkins	Tom Cruise	Brad Pitt	Morgan Freeman	Brad Pitt
Johnny Depp	Brad Pitt	Leonardo DiCaprio	Tom Hanks	Anthony Hopkins	Brad Pitt

FIGURE 17.1 Stimuli used in the pair comparison study of actors. The instructions to the task might read, “On each card, circle the better actor of each pair.”

TABLE 17.1 Similarity (or choice) matrix in a pair comparison study of acting ability ($N = 100$)

<i>Actor</i>	<i>Depp</i>	<i>Cruise</i>	<i>DiCaprio</i>	<i>Freeman</i>	<i>Hopkins</i>	<i>Pitt</i>
Depp	50	60	55	50	70	60
Cruise	40	50	60	70	80	60
DiCaprio	45	40	50	60	60	45
Freeman	50	30	40	50	60	50
Hopkins	30	20	40	40	50	35
Pitt	40	40	55	50	65	50
Mean	42.5	40.0	50.0	53.3	63.3	50.0

DiCaprio and Pitt as comparable, and superior to Cruise and Depp; Hopkins and Freeman were judged the best of the lot, with Hopkins a clear favorite of the sample. If this sample were drawn in such a manner as to make it representative of a population (say, of college students, members of the Motion Pictures Academy of Arts and Sciences, citizens of Monaco, etc.), we might generalize these results to the population from which they were drawn.

If we wanted to develop a comprehensive index of the respondents' choices that had the property of an *interval scale*, we would transform the choice matrix to a proportion matrix, by dividing each frequency by the total number of respondents. Then, following Guilford (1954, pp. 154–177), we would perform a set of statistical operations on the data to determine if respondents' choices satisfied a set of necessary assumptions. The assumptions are focused on the issue of whether the aggregated data accurately reproduce the matrix of data from which they were calculated. If they do, the ordering of the stimuli is said to have equal interval properties. That is, identical differences separating the stimuli are meaningful at all points of the scale. Thus, a .15-unit difference between stimuli falling at one end of the scale is the same as a .15-unit difference between stimuli falling at the opposite end of the scale.

For example, when using the method of pair comparison scaling, if the mean acting abilities of three actors were .60 for Actor A, .50 for Actor B, and .40 for Actor C, then the researcher may be able to state that the extent that Actor A is a better actor over Actor B reflects the same difference in acting ability of Actor B over Actor C. We can infer more from equal interval data than the mere order of the stimuli (as would be the case if an *ordinal scale* were formed).

Data *transitivity* is a central assumption of the scaling operations employed to transform respondents' preferences into a scale of equal interval quality. Transitivity of preferences must be satisfied if the researcher is to convert binary judgments of paired comparison scaling of data that are ordinal quality or higher.⁵ Transitivity implies logical consistency in judgment, such that if a group of respondents feel that Hopkins is a better actor than Pitt, and Pitt better than Cruise, then they should judge Hopkins better than Cruise. Such a set of transitive judgments would be expressed as follows:

If Hopkins is judged better than Pitt,
and Pitt is judged better than Cruise,
then Hopkins *should* be judged better than Cruise.

It sometimes happens that a transitive relationship of this type is not obtained (e.g., in this example, Cruise might be judged better than Hopkins). Intransitive or logically inconsistent choices can be caused by many factors. For example, the stimuli might appear so similar on the choice dimension that respondents cannot differentiate them reliably. Judges who were wildly

enthusiastic about all the actors or hated all of them equally would be hard put to differentiate them in a transitive manner. Another possibility is that the scale along which respondents judge the stimuli is multidimensional. In our actor rating example, it might be that a rater's choice of Cruise over Hopkins was dictated by sex appeal, whereas the choice of Hopkins over Pitt was based on difference in the actors' voice quality; Pitt might be favored over Cruise owing to his appearance in a new movie the respondent enjoyed very much, or by some complex interaction of the voice quality and sex appeal factors (see Tversky, 1969).

Enough unreliable or intransitive choices made by the sample will result in a data set that will not satisfy the minimum criteria of scale quality. That is, tests used to determine the reliability of the scale will indicate that the data upon which the scale is based are not sufficiently trustworthy to put any faith in the resulting scale (Mosteller, 1951). In this situation, the researcher is in a difficult position. The scaling process has failed, but why it failed is uncertain. Some likely possibilities to investigate when seeking the source of the problem involve participants' familiarity with the stimulus dimension, the definitional specificity of the choice dimension, the differentiability of the stimuli, and the dimensionality or clarity of the choice dimension used to judge the stimuli.

Participants' familiarity with the stimuli being judged is the easiest issue to assess, but the hardest to offset. It sometimes happens that some fraction of the participant sample simply is unfamiliar with all the stimuli to be compared. For example, suppose that some of our participants did not know much about Brad Pitt, and could not remember what Anthony Hopkins looked like. They could not be expected to compare these actors reliably with others whom they could recall. Comparisons involving these unfamiliar stimulus objects well might prove unreliable, or intransitive, and thus spoil the overall scale.

Sometimes, the choice dimension is not specified clearly, or, if clearly specified, is not one that participants can use consistently. Returning to our actor example, good acting involves many different qualities. If we are not specific about the particular quality or dimension we wish our sample to use in forming their evaluations, the complexity of the judgment dimension might defeat our attempts at developing a reliable measure. Thus, the researcher must provide clear instructions regarding the specific dimension to be used in judging the stimulus set.

The issue of specificity is related to the multidimensionality of the choice dimension the participants use in forming their judgments (Shiffman et al., 1981). We know that quality of acting can be differentiated along several different dimensions (e.g., quality of voice, dramatic expression, subtlety of body movements, etc.). If our participants were not provided clear and explicit guidelines regarding the dimension to judge the stimuli, and instead used one dimension in one comparison and another in a later comparison, there is a good possibility that the judgments will prove intransitive. The shift in dimensions used from one judgment to another will cause inconsistencies within a respondent. If enough judgments (or judges) suffer from this problem, the scale will not be reliable.

The differentiability of the stimuli also can prove problematic in stimulus scale construction. For example, suppose we wanted to order the stimulus set of popular brands of beer along a dimension of tastiness. We ask our sample to judge among the following seven beers: Amstel, Beck's, Dos Equis, Harp, Heineken, Michelob, and Rolling Rock. Although all our judges might have experienced each brand at one time or another, some of them (not being connoisseurs of the art form) might not be able to distinguish among all, or many of them. This is not necessarily the result of unfamiliarity. For some, it may be that the differences among the various beers are not discernible. In this case, the scaling effort will fail. Indeed, it would fail even if the pair comparison scaling study were performed as a taste test – that is, if participants were given small glasses of beer to taste and compare, rather than trying to match taste with brand

from memory (remember, seven stimuli would require 21 *pair* comparisons, or 42 small glasses of beer). If the beers were relatively indistinguishable – or even if two or three of them were – the researcher might find that the resulting choice matrix would not allow for development of a reliable measurement scale.

Another problem that can arise when we choose to use a pair comparison scaling approach comes about because of participant fatigue. Recall that $N(N-1)/2$ non-repeated pairs can be derived from N stimuli. Thus, if our research problem involved the judgment of 20 actors' ability, or 20 beers, each respondent would have had to make 190 comparisons. Distortion of results attributable to boredom, fatigue, or, in the case of the beer example, intoxication is possible in such situations and would produce a data set that probably would not pass the reliability test. Some researchers reduce the complexity of the problem by not administering the complete set of all possible pairings to the entire sample. That is, respondents judge a different subset of the pair comparisons. Statistical procedures are available to allow this form of pared-down comparative process, but presentation of the complete set of paired stimuli to all judges is preferable. Other researchers divide the task across multiple testing occasions, where half the item pairs are judged first, the remaining subset judged later. This, too, lightens participant load, but variations between test days may introduce error into the process. The problem is best solved by using a relatively restricted number of stimuli which, when paired using all possible combinations, do not produce an overwhelming demand on participants' stamina.⁶

Let's not paint too gloomy a picture, we should recognize the positive features of the pair comparison scaling approach. Under appropriate circumstances, the method can produce an accurate and concise summary of a group's judgments even when the dimension along which the judgments are made has no obvious physical metric. The "appropriate circumstances" involve a set of well-defined stimuli that are discriminable and familiar to the respondents, and a unidimensional judgment rule that is precisely stipulated and followed by all. Further, there should be a reasonable number of stimuli, so the task does not overburden the stamina or cognitive capacities of respondents.

Interval scale results do not allow for absolute judgments, or judgments that entail ratio-level data (e.g., this actor (or beer, or pizza) is twice as good as the competition). Although Anthony Hopkins was rated top in our hypothetical exercise, for instance, it is possible that in an absolute sense, most respondents considered him a very poor actor. Hopkins falling at the top of the scale does not necessarily imply that the respondents thought him a great, or even a good, actor – possibly, they simply found him less bad than the others in the comparison group. Only data of *ratio* quality provide a true zero-point, which in the present instance would allow us to determine whether the aggregate participant sample viewed the top-rated actor as good or bad, rather than better than the competition. If a true zero-point had been established (by other scaling methods), we could infer actors falling above zero had been viewed positively by the sample; those below it were seen as bad actors. In addition, the presence of the true zero would allow us to determine the degree to which one stimulus exceeded another on the dimension used by the sample to differentiate the stimuli.⁷

The Method of Rank Order

In the method of **rank order scaling**, a set of stimuli is examined by having each respondent order all stimuli simultaneously along a choice dimension stipulated by the researcher. Rank order scaling is a comparative stimulus scaling technique that avoids many of the problems inherent in paired comparison scaling, while producing results that closely approximate those of the more laborious method of comparing all possible pairs of stimuli (Misra & Dutt, 1965;

Nishisato, 1978). Modifying the previous example, we could present our six actors to participants, and ask them to rank them in order of acting ability, with 1 representing the best, and 6 the worst, actor. If using index cards, the names of the six actors are individually written on different index cards. Then, participants merely sort the names from best to worst actor. From the data obtained in this simple operation, we could construct an equal interval scale of participants' judgments (see Guilford, 1954, pp. 178–196 for a description of the statistical operations used).

Notice that this ranking technique avoids two potentially serious problems of the pair comparison scaling approach. First, because all objects of judgment are presented at once, and the respondent simultaneously ranks them, judgmental intransitivity is impossible. If A is ranked over B, and B over C, then A must be ranked over C in the rank order approach. In addition, the method of rank order avoids some of the administrative drudgery due to paired comparisons, especially when large numbers of stimuli are to be compared. Remember, 190 pair comparisons would be generated from a set of 20 stimuli. If the method of rank order were used, the task would be confined to the judgments involved in ordering only those 20. Thus, the technique appears to demand less of the respondents. Nonetheless, some prefer the pair comparison scaling approach in research involving small numbers of stimuli (say, ten or fewer). In these studies, participants' responses are thought to be more reliable than in a rank order task because only two objects of judgment are involved in any given pair comparison (e.g., Crano & Cooper, 1973). Proponents of pair comparison scaling believe that the dimension along which stimuli are to be judged can be held in mind more faithfully when it need be applied in choices involving only two stimuli. In the rank order method, it could be argued that essentially all the stimuli are judged at once. On the other hand, if a large number of items (more than 20) are in the stimulus set, even the rank order method may be overly demanding for respondents who may find it difficult to hold that many stimuli in mind at one time.

Aside from these distinctions, the methods of pair comparison and rank order share noteworthy similarities. First, both methods are used to generate *comparative judgments* of a set of stimuli. Either scaling process, that is, provides information regarding relative differences between stimuli as judged by the participant sample, and these differences are arranged on a scale of equal intervals. Note that neither method provides information regarding the judges' *absolute* appraisal of the rated objects on a ratio scale.

Another important similarity between these two stimulus scaling methods concerns the assumptions regarding differences between participants' ratings. Because responses to stimulus objects are pooled across participants in both methods, differences between participants are considered irrelevant and therefore are ignored. Respondents are viewed as replicates, and differences among them are attributed to error or unreliability. In other words, by virtue of the way in which the data are assembled and combined, these approaches assume that all participants in a sample would produce the identical pattern of choices if their judgments were perfectly reliable. This assumption is necessary to justify the pooling of responses over participants. Fortunately, violations of this assumption can be tested (see Guilford, 1954; Torgerson, 1958) and if the violation is not too extreme, it can be offset.

Multidimensional Scaling Models

The scaling techniques we have discussed to this point are designed to develop unidimensional scales. Indeed, researchers working in this scaling tradition generally strive to achieve unidimensionality in their scales. If judges can slide from one dimension to another when comparing stimuli, the classic psychometric methods fail to produce reliable measures, and this is especially problematic for the method of pair comparison scaling. Some investigators, however,

have argued that unidimensional scales do not adequately reflect the complexity commonly encountered in our everyday lives. For example, we can judge actors in terms of “acting ability,” especially when that term is strictly defined for us. However, the single dimension we forced participants to use in our study to judge the goodness or badness of acting ability probably grossly oversimplifies the complex judgment scheme that people use in reality. Scott (1968) recognized this fact when he argued that using a single dimension to construe a complex stimulus object (an actor, a beer, etc.) “is patently unrealistic if one takes seriously the widely held psychological principle that any response is multiply determined” (p. 250). Arguments of this nature have proved persuasive to many, and multidimensional scaling approaches have become more widely used in the social sciences. An extended discussion of multidimensional scaling is beyond the scope of this book, and it is fair to say that multidimensional scaling approaches are not nearly as well developed as might be expected. Recall that Scott made his observations more than 50 years ago. However, some general points, along with the preceding discussion of unidimensional approaches, should prove a useful introduction to these techniques.

Unfolding

Coombs (1964) supplied an appropriate point of transition between unidimensional and multidimensional scaling, and between stimulus and individual difference scaling. In Coombs’ **unfolding technique**, the preference order of a set of stimulus objects as well as a set of respondents is determined simultaneously along the same dimension stipulated by the researcher. In this sense, respondents are considered an additional set of “stimuli” to be ordered along a dimension. Coombs’ unfolding technique thus bridges the two scaling traditions – stimulus scaling and individual difference scaling (Coombs et al., 1970).

Consider the following example: Sam, Jenny, and Brian are at a point in their college careers when they must choose an academic major. For convenience, let us assume that these three students can choose among six different areas, and the rank ordering of their choices is as presented in Table 17.2.

Although their choices obviously are quite distinct, a single preference structure can be constructed that summarizes the inclinations of all three students toward each of the six possible academic majors, as ranked from first choice to sixth choice, depicted in Figure 17.2. To determine whether the scale of Figure 17.2 adequately summarizes the ordered choices of our three students, consider the point at which Brian falls on the preference scale. In terms of scale distances, the major closest to Brian is theatre, then psychology, then chemistry, etc. Thus, Brian’s placement on the scale accurately reproduces his preference rankings of Table 17.2.

We could create a more graphic illustration of Brian’s preferences by vertically “folding” the scale at his ideal point (i.e., the point at which he intersects the scale), with distances closer to

TABLE 17.2 Three students and their ranked preferences on six academic majors

<i>Ranked Preference</i>	<i>Jenny</i>	<i>Sam</i>	<i>Brian</i>
1	Chemistry	English literature	Theatre
2	Physics	Psychology	Psychology
3	Theatre	Art	Chemistry
4	Psychology	Theatre	English literature
5	English literature	Chemistry	Physics
6	Art	Physics	Art

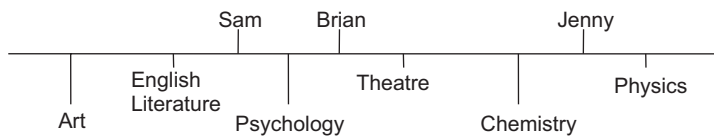


FIGURE 17.2 Underlying preference structure for three students and six majors.

this fold preferred over majors that are farther away – hence the term “unfolding technique” to describe Coombs’ approach. Using distances from each student’s ideal point as indicators of relative preferences, we can reconstruct exactly the preference rankings of each of the students, as they were presented in Table 17.2. If one is seeking to examine individual differences, each student’s position in the top half of Figure 17.2 should be contrasted: Sam and Brian, being more proximal, are more similar in their ranking sequence for majors than the rankings made by Jenny. Inspection of the location of the majors in the lower half of Figure 17.2 informs us that the majors of Chemistry and Physics, which are closer in distance, tend to be more similar in the ordering sequence than to the majors of Art or English Literature. Thus, in the unfolding technique, scaled data allow for the simultaneous evaluation of stimuli differences and individual differences. As Galanter (1966) observed, “by using this (unfolding) technique, we see...that although different people may exhibit different preferences...it may very well be the case that the differences in the revealed preferences conceal an underlying consistency in the preference structure” (p. 131).

The utility of the unfolding technique is enhanced if the dimension along which the preference stimuli are ordered is somehow identifiable. In the present example, we could hypothesize that the dimension that helped determine the ways the students arranged their preferences was the degree to which mathematical ability was required by each of the academic majors. If the school records of our three students showed that Jenny’s mathematical aptitude test score exceeded that of Brian’s, whose score was greater than Sam’s, then our confidence in our description of the critical dimension would be enhanced, as would the utility of the scale.

A good research example of the use of the unfolding technique was provided by Poole (1981), who investigated the ratings members of the U.S. Senate received from 26 special interest groups, which ranged from Americans for Democratic Action on the left to the National Taxpayers’ Union on the right. Poole found that a single dimension, on which Senators were arranged in terms of their liberal or conservative leanings, accounted for nearly 80% of the variance in the special interest groups’ ratings. Furthermore, Senators’ votes on many crucial issues could be predicted based on their relative standing on the liberal-conservative dimension that Poole derived. And, the accuracy of these predictions surpassed that based on political party affiliation. It is possible that the students’ (or organizations’) preferences could have been so diverse that a simple one-dimensional preference structure could not have been constructed in such a way that it accurately summarized all choices. For example, suppose that mathematical aptitude and verbal aptitude were the two factors that influenced each of the student’s choices differently. In this case, the unfolding solution would have required two dimensions, and the students’ ideal preference points would be located not on a single scale, but somewhere in the two-dimensional space described by mathematical aptitude and verbal aptitude.

Multidimensional Scaling Approaches

Many multidimensional scaling approaches have been developed over the years (e.g., see Carroll & Arabie, 1980; Fox & Escue, 2021; Kruskal, 1964a, 1964b; Kruskal et al., 1977; Shepard,

1962a, 1962b; Shepard et al., 1971). Usually, **multidimensional scaling** (MDS) relies on ratings of the degree of similarity or dissimilarity among pairs of stimuli, rather than binary judgments or preference ordering to identify the number and types of dimensions that underlie a set of stimuli. In research designed to determine the factors or dimensions that underlie the raw data, factor analysis is appropriate for individual differences scaling, but MDS is used for scaling comparative ratings. Specifically, MDS requires the sample to assign similarity ratings for all possible pairs of a set of stimuli. Data in this format are not suitable for factor analysis.

In an interesting example of the use of multidimensional scaling, Wish et al. (1970) identified the number and types of dimensions people used when judging the similarity of various nations. To accomplish this, Wish and colleagues drew up a list of 21 nations, and presented all possible pairs of these nations to each of 75 respondents, who were instructed to judge their degree of similarity (using a 9-point judgment scale). The degree of similarity between each pair of nations was calculated across respondents, and used to construct a similarity matrix, which would resemble that of the actor example presented in Table 17.1. Multidimensional analysis of these similarity judgments revealed that four dimensions largely determined respondents' judgments of nations: political ideology, level of economic development, geography and population, and the culture and predominant race of the countries. Although the study simply instructed the sample to rate the degree of similarity for each nation pair, the MDS analysis indicated participants tended to apply these four dimensions when making their judgments of similarity. It seems likely that most respondents were not completely aware that these dimensions had influenced their similarity-dissimilarity judgments of nations.

Earlier MDS research that instructed participants to rate the similarity of personality traits (stimuli) revealed two dimensions: intelligence and sociality (Rosenberg et al., 1968). For example, the traits of "reliable" and "practical" were each perceived as high in both intellectual and sociality, but the traits of "boring" and "dishonesty" were each perceived as low in both features.

These dimensions – intelligence and sociality – are synonyms for competence and warmth, respectively, two features used in the development of the *stereotype content model* (Fiske et al., 2006). Although these latter researchers did not use pair comparison scaling to develop their central theoretical dimensions, MDS might be used to extend their line of research. For example, research using MDS may be conducted to evaluate the perceived similarity/dissimilarity of social groups. Suppose participants are presented with pairs of words depicting social groups (e.g., housewife and wealthy), and instructed to assign a rating of 1 (extremely dissimilar) to 10 (extremely similar) to each pair. Analysis of the similarity ratings might reveal the emergence of competence and warmth, two dimensions that might underlie respondents' judgments, and which appear to Fiske and associates as intelligence and sociality.

Presented in Figure 17.3 are possible results of the *perceptual map* that might be found in a study of this type. Noting where the data points are situated, the illustration shows that northerners, lawyers, and wealthy people are perceived to be relatively high in competence; housewives, elderly, blind, and the homeless are low on the dimension. Furthermore, housewives, elderly, blind, southerners, and northerners are perceived high in warmth, but the homeless, lawyers, and the wealthy people are not. Taking into consideration both dimensions, northerners are viewed as high in competence and warmth, but the homeless are considered low on both dimensions. Research on the stereotype content model suggests that when forming social impressions of others, we tend to classify them on these two stimulus dimensions.

Alvaro and Crano (1997) used MDS to estimate the proximity of a set of attitudinal beliefs held by their participant sample. They presented a set of attitude items, and asked participants to estimate the likelihood that if they changed their attitude about one of the items, they would

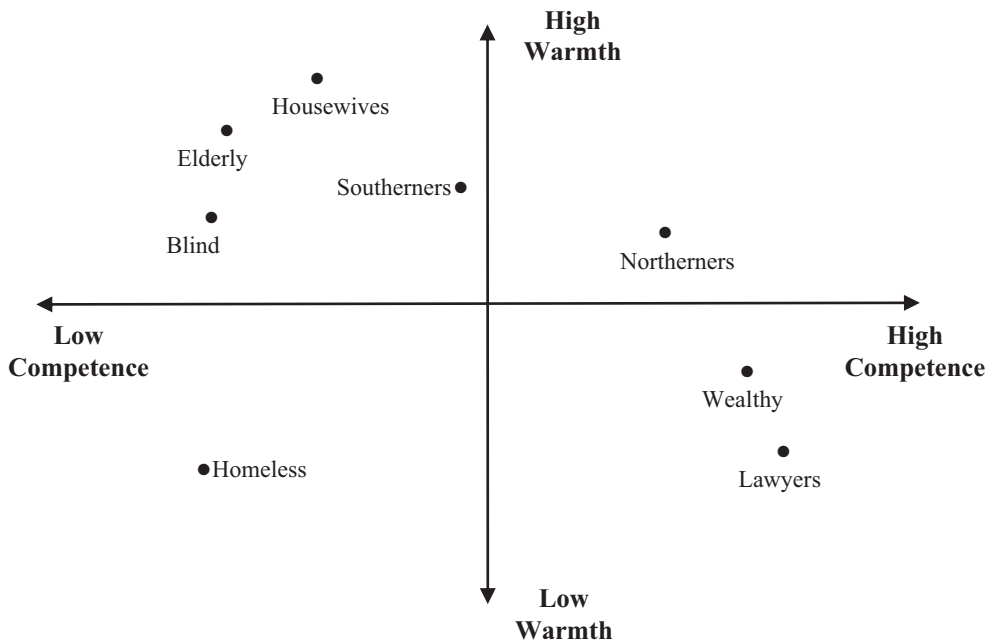


FIGURE 17.3 Multidimensional scaling on dimensions of competence and warmth.

change their attitude on another. Following is an example of the format they used to investigate perceived similarities and difference among any pair of items.

If you changed your mind regarding your position on HOMOSEXUALS IN THE MILITARY, what is the percentage likelihood that you would also change your position on ABORTION?

Percentage = __ (Note: percentages may range from 0–100%)

A higher percentage score indicated greater similarity between the two items. Participants' responses were used to create a matrix of similarities among all possible pairs of items. The MDS approach produces a map, or a picture, of the spatial relationships among all the stimuli that are presented. The MDS analysis of Alvaro and Crano (1997) suggested the emergence of two dimensions of attitude objects in the similarity matrix.⁸ For purposes of testing their theory of attitude change, the researchers were searching for highly proximal attitude objects. So, as expected, the abortion item mapped very closely to the contraception item. Participants thought that if they were to change their attitude toward abortion, they also would change their attitude toward contraception. However, the MDS also revealed proximal attitude objects that participants did not think were related and these were used in the design of a follow-up study that investigated ways in which a persuasive message about a specific attitude object could affect a seemingly unrelated attitude object (see also Crano & Chen, 1998).

The MDS approach is a highly useful, if underused method in social research. Although underused, the approach has not been ignored completely. Some of the more intriguing uses of MDS in the social science literature involve studies of the ways in which Alzheimer's patients cluster common objects (e.g., instruments, animals, etc.), as compared with the mappings of elderly non-Alzheimer's patients (Ober & Shenaut, 1999), differences related to variations in

music preferences (Tekman, 1998), adolescent functioning (Wardenaar et al., 2013), educational research (Ding, 2018), risk behaviors (Dong & Ding, 2012), and factors that affect judgments of emotion in faces (Halberstadt & Niedenthal, 1997).

Conclusion

We consider the underuse of stimulus scaling methods a missed opportunity. The techniques are based on classic psychophysical methods developed over many years and have proved robust and reliable. They provide relatively fine-grained information on dimensions underlying beliefs and preferences and reveal how similar or dissimilar these items are on the identified dimensions. Recall that variations among participants are assumed to be a function of error (or noise) in stimulus scaling. If sufficient disagreement exists, the stimulus scaling procedures will fail; that is, their statistical analysis will suggest that the summary “group attitude scale” is not trustworthy, but most often the approaches produce reliable summary information of interval-level quality. We hope this brief discussion of some of the more widely known techniques will encourage their more extensive use in social research.

Notes

- 1 This combination of judgments across all respondents requires the earlier noted assumption that respondents would all judge the stimuli identically were it not for measurement error. If this assumption were not made, it would not be logical to combine data in this way.
- 2 Some of the “crimes” Thurstone used in his study (e.g., abortion, adultery) would not be considered illegal in some U.S. jurisdictions today.
- 3 Borg (1988, p. 60) suggested that participants in Thurstone’s (1927) sample might have read “seduction of a minor” as referring to “having sex with a girl under 21,” whereas today’s students might have read the same item as “molestation of a three-year-old.”
- 4 To calculate the number of pairs that will result from a given number of stimuli, use the following formula: $p = [n(n - 1)]/2$, where p = number of pairs, and n = the number of stimuli. Thus, in our acting quality study, we would produce $[6 \times 5]/2$ pairs. Thurstone (1927) used 19 different crimes, and thus produced 171 pairs of stimuli.
- 5 Remember in Chapter 1 we stated that interval level measures require the properties of both ordered sequence and equal differences between measurement points on the construct.
- 6 Alternatively, pictorial approaches in which participants arrange all objects in two-dimensional space simultaneously also are available, but they will not be considered here (see Hout et al., 2013).
- 7 True zero refers to the fact that the property is completely absent (e.g., I lost 0.0 lbs on my diet, or I have \$0.00 in my wallet, etc.). Unlike physical constructs such as weight, “acting ability” is a psychological construct that does not really have a true zero-point (one can be a good, mediocre, or poor actor, but it does not make sense to talk of zero-point of acting ability, or for an actor to have absolutely no acting ability – though we all might have witnessed performances in which this possibility came to mind).
- 8 In MDS, if the mapping process is successful, the distances derived from the map will equal those found in the similarity matrix. Increasing the number of dimensions usually results in a better fit of map with matrix, but the trick is to scale the similarity matrix using as few dimensions as possible.

References

- Alvaro, E. M., & Crano, W. D. (1997). Indirect minority influence: Evidence for leniency in source evaluation and counter-argumentation. *Journal of Personality and Social Psychology*, 72, 949–965.
- Borg, I. (1988). Revisiting Thurstone’s and Coombs’ scales on the seriousness of crimes. *European Journal of Social Psychology*, 18, 53–61.
- Carroll, J. B., & Arabie, P. (1980). Multi-dimensional scaling. *Annual Review of Psychology*, 31, 607–649.
- Coombs, C. (1964). *A Theory of Data*. New York: New York, NY: Wiley.
- Coombs, C. H. (1967). Thurstone’s measurement of social values revisited forty years later. *Journal of Personality and Social Psychology*, 6, 85–91.

- Coombs, C. H., Dawes, R. M., & Tversky, A. (1970). *Mathematical psychology*. Englewood Cliffs, NJ: Prentice-Hall.
- Crano, W. D., & Chen, X. (1998). The leniency contract and persistence of majority and minority influence. *Journal of Personality and Social Psychology*, 74, 1437–1450.
- Crano, W. D., & Cooper, R. E. (1973). Examination of Newcomb's extension of structural balance theory. *Journal of Personality and Social Psychology*, 27, 344–353.
- Dawes, R. M., & Smith, T. L. (1985). Attitude and opinion measurement. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (pp. 509–566). Reading, MA: Addison-Wesley.
- Ding, C. S. (2018). *Fundamentals of applied multidimensional scaling for educational and psychological research*. Cham: Springer International Publishing AG.
- Dong, Y., & Ding, C. S. (2012). Adolescent risk behaviors: Studying typical and atypical individuals via multidimensional scaling profile analysis. *Journal of Adolescence*, 35, 197–205.
- Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2006). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, 11, 77–83.
- Fox, B., & Escue, M. (2021). Evaluating and comparing profiles of burglaries developed using three statistical classification techniques: Cluster analysis, multidimensional scaling, and latent class analysis. *Psychology, Crime & Law*, 1, 34–58.
- Galanter, E. (1966). *Textbook of elementary psychology*. San Francisco, CA: Holden-Day.
- Guilford, J. P. (1954). *Psychometric methods*. New York: McGraw-Hill.
- Halberstadt, J. B., & Niedenthal, P. M. (1997). Emotional state and the use of stimulus dimensions in judgment. *Journal of Personality and Social Psychology*, 72, 1017–1033.
- Hout, M. C., Goldinger, S. D., & Ferguson, R. W. (2013). The versatility of SpAM: A fast, efficient, spatial method of data collection for multidimensional scaling. *Journal of Experimental Psychology, General*, 142, 256–281.
- Kruskall, J. B. (1964a). Multidimensional scaling: A numerical method. *Psychometrika*, 29, 1–27.
- Kruskall, J. B. (1964b). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29, 115–129.
- Kruskall, J. B., Young, F. W., & Seery, J. B. (1977). *How to use KYST 2: A very flexible program to do multidimensional scaling and unfolding*. Murray Hill, NJ: Bell Telephone Labs.
- Misra, R. K., & Dutt, P. K. (1965). A comparative study of psychological scaling methods. *Journal of Psychological Research*, 9, 31–34.
- Mosteller, F. (1951). Remarks on the method of paired comparisons: III. A test of significance for paired comparisons when equal standard deviations and equal correlations are assumed. *Psychometrika*, 16, 207–218.
- Nishisato, S. (1978). Optimal scaling of paired comparison and rank order data: An alternative to Guttman's formulation. *Psychometrika*, 43(2), 263–271.
- Ober, B. A., & Shenaut, G. K. (1999). Well-organized conceptual domains in Alzheimer's disease. *Journal of the International Neuropsychological Society*, 5, 676–684.
- Poole, K. T. (1981). Dimensions of interest group evaluation of the U.S. Senate, 1969–1978. *American Journal of Political Science*, 25, 41–54.
- Rosenberg, S., Nelson, C., & Vivekananthan, P. S. (1968). A multidimensional approach to the structure of personality impressions. *Journal of Personality and Social Psychology*, 9, 283–294.
- Scott, W. A. (1968). Attitude measurement. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology*, Vol. 2. *Research methods* (2nd ed., pp. 204–273). Reading, MA: Addison-Wesley.
- Shepard, R. N. (1962a). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, 27, 125–140.
- Shepard, R. N. (1962b). The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika*, 27, 219–246.
- Shepard, R. N., Romney, A. K., & Nerlove, S. (Eds.) (1971). *Multidimensional scaling: Theory and applications in the behavioral sciences*. New York: Academic Press.
- Shiffman, S. S., Reynolds, M. L., & Young, F. W. (1981). *Introduction to multidimensional scaling: Theory, methods, and applications*. Bingley, UK: Emerald Group Publishing.
- Tekman, H. G. (1998). A multidimensional study of music preference judgments for excerpts of music. *Psychological Reports*, 82, 851–860.

- Thurstone, L. L. (1927). Method of paired comparisons for social values. *Journal of Abnormal and Social Psychology*, 21, 384–400.
- Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.
- Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, 76, 31–48.
- Wardenaar, K. J., Wigman, J. T. W., Lin, A., Killackey, E., Collip, D., Wood, S. J., ... & Yung, A. R. (2013). Development and validation of a new measure of everyday adolescent functioning: The Multi-dimensional Adolescent Functioning Scale. *Journal of Adolescent Health*, 52, 195–200.
- Wish, M., Deutsch, M., & Biener, L. (1970). Differences in conceptual structures of nations: An exploratory study. *Journal of Personality and Social Psychology*, 16, 361–373.

18

INDIRECT AND IMPLICIT MEASURES OF COGNITION AND AFFECT

As in any field of science, the use of the scientific method in social research requires that our assertions be based on observable phenomena (see Chapter 1). Whatever inferences we wish to make about the causes and processes underlying social behavior must first be grounded in observations that can be recorded and replicated. Research that employs the experimental method involves manipulating some aspect of the physical or social environment, and then observing and recording some type of response of participants as the dependent variable. The observed response could be an overt behavior or action of some kind (e.g., stopping to give help, pressing a button to deliver an electric shock to another person, choosing a gift). More often, however, the observed response is a written or oral report from a participant of his or her reactions to the situation, a judgment, or a decision. Similarly, in survey research involving interviews or questionnaires, the observations consist of respondents' self-reports of their behaviors, feelings, or beliefs. Because inner experiences – personal feelings and mental life – are not directly observable, social researchers often must rely on people's introspective reports of their private experiences to acquire data that are amenable to measurement and quantification.

In previous chapters we raised several issues and problems that must be considered in evaluating the validity of self-report measures as accurate assessments of respondents' true feelings and beliefs. When respondents are aware that they are participants in a scientific investigation, evaluation apprehension and social desirability concerns may lead them to adjust their responses to meet personal or social standards or expectations (see Chapter 7). Participants may be *unwilling* to report on their true feelings or reactions, particularly when embarrassing, sensitive, or politically charged issues are at stake. Even when respondents are willing to provide truthful and candid accounts, they may be *unable* to report accurately on their own inner feelings or mental states. In their classic research, Nisbett and Wilson (1977) argued and documented evidence showing that individuals do not always have conscious access to many of the mental processes that underlie their behaviors or decisions, at least not in a manner that they can verbalize and articulate.

Given the evidence that respondents often are either unwilling or unable to provide valid reports on certain aspects of their inner experiences or mental processes, exclusive reliance on introspective self-reports as our principal source of information about internal reactions is problematic. Fortunately, over the past few decades, social researchers have developed an armory of

new techniques and procedures for tapping the “inner world” of cognitive processes and affective experiences that do not rely on conscious self-report. The techniques that we describe in this chapter cannot fully replace self-report measures as a mainstay of social research, but they can augment more traditional techniques by providing different types of information that are not susceptible to the same motivational or capability limitations.

Indirect Measures

The rating scale approaches discussed in Chapter 16 involve the *direct assessment* of respondents’ expressed attitudes, preferences, judgments, knowledge, etc., because respondents are asked explicitly about these feelings or beliefs. Over the years, many techniques focused on the *indirect* assessment of attitudes have been developed. In indirect attitude assessment, “the investigator interprets the responses in terms of dimensions and categories different from those held in mind by the respondent while answering” (Kidder & Campbell, 1970, p. 336). Indirect approaches are used to reduce possible distortions that might come about when respondents answer questions in a socially desirable or normative manner to place themselves in a more favorable light. Many researchers feel that they can obtain more accurate assessments by having respondents focus their attention on irrelevant but compelling features of the task. They hope that by using misdirection or assessing reactions of which people are unaware, their respondents will lower their defenses and present a more valid picture of their attitudes, beliefs, or judgments.

Many indirect techniques have been developed, with varying degrees of success (Dovidio & Fazio, 1992). For example, suppose a researcher was interested in indirectly assessing respondents’ attitudes toward labor unions, and decided to employ a **sentence completion task**, an indirect measurement in which a partial sentence is presented, and participants are required to complete the sentence. In the instructions, the investigator might ask respondents to be “as creative as possible in completing the following sentence stems,” among which would be the following:

The cost of living...
The labor unions have...
Labor unions are...

Participants are to indicate the word or phrase that first comes to mind to complete each sentence. A presumption of the sentence completion approach is that the information most at the tip of the tongue or accessible in the mind of the participant will be offered as the response. If given the sentence stem of “My research methods course is ...” students who answer “totally awesome” or “making my life complete” probably are enjoying the material more than students responding with “making me sweat, and not in a good way” or “preventing me from having an active social life.” “Filler” items not of interest to the researcher often are included in such tasks to mask the intent of the study.

Other approaches to assess attitudes surreptitiously include the **thematic apperception test**, an indirect measurement in which the respondent views a deliberately ambiguous picture and then generates a story about the characters and scenario depicted in the picture. For instance, a drawing may depict two youths holding hands in a playground, but the drawing is sketched such that participants will have to interpret the gender of each child, and whether they are holding hands because they are good friends, have a “crush” on each other, or are trying to avoid being punched by the other. The content of the respondent’s story can be analyzed in

terms of how the image is interpreted and the aspects of the image that most strongly drew the participant's focus. The intent of the measure is far from transparent, and thus, the hope is that the method will allow the researcher to obtain more honest and unbiased answers than would more direct assessment methods.

It must be acknowledged that these forms of assessment make considerable demands on the time and technical expertise of the investigator because the responses typically gathered in these methods do not lend themselves to easy scoring. Generally, coders must be trained to perform systematic content analyses of respondents' qualitative responses, even when more automated analytic applications are used (Chapter 14). This training is labor intensive and time consuming. In addition, it is necessary that some estimates of the interrater reliability of the scoring procedure be developed (Chapter 13). If the interrater reliability level is unacceptable, the coding system must be revised or the coders retrained.

A briefer and less demanding version of indirect assessment is the *word completion task*. In this method, respondents are shown a series of incomplete words, or word fragments, and asked to fill in the missing letters with the first word that comes to mind. For example, the test items might be:

c_ff_ _;
m_ _ _ _er;
de_ _ _

If a respondent completes these fragments as “coffin” (rather than “coffee”), “murder” (rather than “mother”), and “death” (rather than “debit”), we might assume that they have mortality on their mind.

Information Processing: Attention and Memory

Other methodologies for indirect assessment of attitudes and preferences rely on **implicit responses**, responses to stimuli that participants are not necessarily aware that they are making. Because many of these techniques for “getting inside the head” (Taylor & Fiske, 1981) have been adapted from procedures developed by cognitive scientists, they are often referred to collectively as *social cognition* methodologies (Hamilton & Stroessner, 2021), although many are intended to assess affect, emotions, and motives as well as cognitive processes.

Many of the techniques for assessing cognition derive from a general model of information processing that assumes that knowledge about the world and experiences is acquired and remembered through four stages or operations: *attention* (what information is attended to), *encoding* (how that information is understood and interpreted at the time of intake), *storage* (how information is retained in memory), and *retrieval* (what information is subsequently accessible in memory). Methodologically, these processes may be measured or documented by various techniques.

Measures of Attention

The information processing model assumes that attention is a limited resource that is selectively distributed among the myriad of visual, auditory, and other sensory stimuli that bombard us at any point in time. Attention serves as a selective filter of the voluminous information that comes to us from the external world: People must first pay attention to information before it can be passed on to subsequent stages of memory. It is further assumed that the specific stimuli

that capture and hold a person's attention are those that are most salient or important to the perceiver at the time. Thus, by measuring which inputs a person attends to when multiple stimuli are available, or measuring how long the person attends to specific stimuli compared to others, researchers have an indirect way of assessing what is important, interesting, or salient to that individual.

Visual attention. The majority of research on attention (and hence, methods for assessing attention) focuses on the perception of visual information – either of actual events or displays of pictures, words, or other symbols. Measures of visual attention involve tracking direction and duration of eye gaze, i.e., when and how long the perceiver's eyes are fixated on a particular object or event presented in the visual field.¹ Atalay et al. (2012), for example, assessed visual fixation patterns to analyze consumers' choices in horizontal displays of merchandise. Memory-based methods failed to relate to choices, but eye gaze did. In another application, McArthur and Ginsberg (1981) used eye tracking to measure selective attention to specific individuals when they were forming impressions about other people.

Precise measurement of eye fixation patterns entails heavy technology and may require that participants hold their heads still in an apparatus so that their eye movements may be tracked and recorded across small distances. Computerized methods for producing visual displays and recording sequential eye movements are routinely used in visual research but are generally less accessible or useful for social research. For purposes of social information processing, video recording participants' faces during social interactions or decision-making will usually provide sufficient information about location and duration of eye gaze to determine what is being attended to. Olson and Zanna (1979), for example, recorded participants' eye gaze while they inspected a pair of painting reproductions in a study of selective attention to preferred versus non-preferred objects. In this case, the measure of interest was how much time participants spent looking at the painting they had chosen compared to the alternative that was not chosen. Eye gaze proved to be a sensitive measure of selective allocation of visual attention and preference.

Interference as a measure of attention. Another way of assessing how much attention is being devoted to a particular stimulus is to determine whether the presence of a distracting stimulus interferes with attending to or processing of a target stimulus. The well-known "Stroop effect" (Stroop, 1935) is an example of an interference-based measure of unintended attention. In this research paradigm, participants are presented with a series of words printed in various colors and are instructed to not read the word but to identify aloud the color of ink in which each is printed. The relation between the word and the ink color is varied. In *congruent trials*, each written word corresponds to the ink color (e.g., the word "green" printed in green ink), so identification of ink color should be facilitated. In *incongruent trials*, the written word itself is a color different from the printed ink color (e.g., the word "red" printed in green ink), so the semantic meaning of the word interferes with the participant's ability to identify the ink color correctly. Thus, it takes longer to name the ink color on incongruent trials compared with trials in which the word is congruent with the ink color. This response interference is an indication that attending to the semantic meaning of the word is automatic and cannot be suppressed even when instructed to identify the ink color instead. The Stroop effect can involve a between-group design in which participants are randomly assigned to one of the two conditions, or a within-subjects counterbalanced design in which half the participants receive the congruent condition first and the other half receive the incongruent condition first.

The Stroop effect has been adapted as a general measure of automatic attention to semantic content (see Logan, 1980, or MacLeod, 1991, for reviews). For example, Pratto and John (1991) used the effect to study automatic attention to negative words. Respondents in this experiment were to identify the colors in which words of personality traits were written. Some of the words

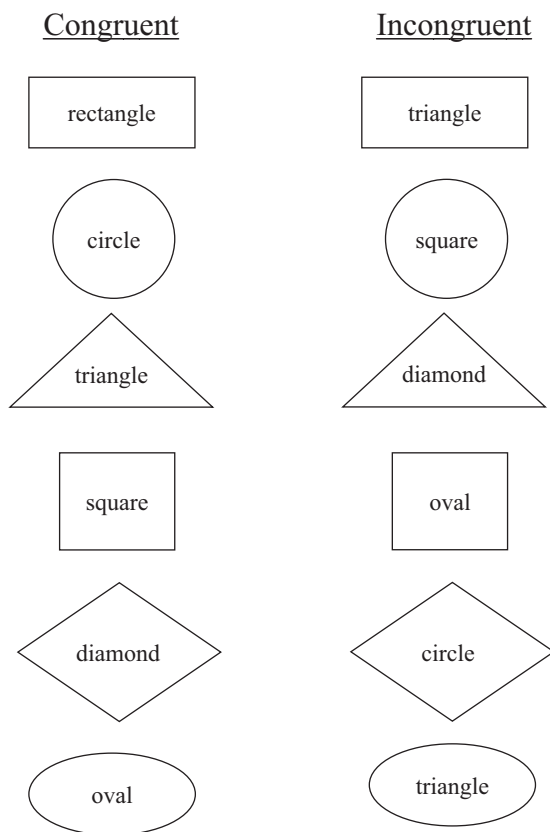


FIGURE 18.1 Stroop task using shape-word congruent and incongruent conditions.

referred to desirable traits (such as “honest”) and others to undesirable, negative traits (e.g., “sadistic”). The delay in reaction time (or latency) to respond by stating the correct ink color was consistently longer when undesirable words were presented, suggesting that participants had difficulty ignoring the distracting effect of social stimuli with strong negative connotations.

Presented in Figure 18.1 is an illustration of a variant of the Stroop experiment, with the congruent and incongruent conditions on the left and right columns, respectively. Starting with one of the conditions, begin at the top and verbally state each shape aloud (rather than state the word) as fast as you can. If you make a mistake, try again until the shape is correctly stated. Proceed until you have reached the bottom. Afterward, verbally state each shape for the other condition. In the congruent condition, where there was correspondence between shapes and words, these shapes were likely correctly stated at a relatively rapid rate. In the incongruent condition, where the shapes did not correspond with their semantic meanings, it should take longer to name the shapes. The delayed reaction time in the incongruent condition occurs due to reading – a process we have practiced all our lives and therefore has become relatively automatic – interfering with the ability to name the shapes. Essentially, the conflicting processes of reading and shape identification simultaneously competed for our attention. The additional amount of time it requires to complete the incongruent condition over the congruent condition indicates the extent of interference.

Processing time. A third method for assessing how important or significant specific information is to the perceiver is to measure the amount of time the individual spends viewing or

contemplating the information before making a decision or judgment or moving on to another processing task. Duration of eye gaze is one measure of processing time, but more often this method is used when information is presented sequentially, as, for example, images displayed successively, or successive photographs or text projected on a computer screen. If the participant is given control over the movement from one screen to the next, the amount of time spent viewing a particular screen provides a measure of processing time for that information.

Processing time is a somewhat ambiguous measure of attention or interest because it combines time spent encoding or interpreting the stimulus object (the second stage of information processing) with simply attending to it (first stage). Thus, the administration of more complex or ambiguous stimuli may engage longer processing time independent of their interest value. Nonetheless, the relative amount of time spent viewing each piece of information in a sequence often may provide useful information about the types of information that attract the most attention and cognitive effort. For instance, Fiske (1980) found that social stimuli (slides depicting social behaviors) that were rare, extreme, or negative elicited increased processing time. Similarly, Brewer et al. (1981) used looking time to demonstrate that information that is inconsistent with category stereotypes takes longer to process than information that is consistent with stereotypic expectancies.

Measures of Memory

The extent to which individuals can remember information or experiences provides significant clues regarding which information has been encoded and how it is stored in memory. In most social research employing memory as a dependent variable, the memory “test” is introduced unexpectedly and without advance warning to assess what is encoded and stored spontaneously when perceivers are not deliberately attempting to memorize information. Memory of stored information may be accessed in either of two ways: recall or recognition. In a **recall measure**, the participant reports what he or she remembers about information that was presented previously. This task requires that the respondent first search and retrieve relevant material from memory, and then make the decision of whether the retrieved information is correct (i.e., that the person believes it was present on the occasion being recalled). Bypassing the search and retrieval stage, in a **recognition measure** the participant simply decides whether the material currently shown matches his or her memory of what was previously presented. The recognition process is considered less cognitively effortful and faster than the recall process, as recognition requires only familiarity with the previously exposed material.

An example of a test of recall is the open-ended question, “What is the capital of California?” In attempting to answer the question, you would search your memory stores to retrieve this information and then decide if what you have in mind is the correct answer. A test of recognition is exemplified by a multiple-choice question, for example, “Which is the capital of California? (a) San Diego, (b) Los Angeles, (c) Sacramento, (d) Fresno.” Having a sense of familiarity with this knowledge in the context of available options is sufficient for you to select the correct response. Recognition tasks are easier because it is only necessary to decide whether one of the presented options correctly matches what was stored previously in memory. Obviously, if you had never been exposed to the fact that the capital of California is Sacramento, this information would not be stored in memory, and recall and recognition responses would reflect random guesses.

Recall measures. The typical paradigm for recall memory experiments involves an initial presentation of stimulus information, which may be anything from a list of words, a series of pictures, a written description or story, or a videotaped event. Usually, the participant is given

some cover story about why he or she is viewing the information (e.g., evaluating the writing style, forming an impression) that does not involve explicit instructions to remember the material being presented. After the presentation, some time is allowed to lapse, and then the participant is asked to report what they can remember of what was presented earlier.² With *free recall* tasks, participants are given no further instructions about what to search for in memory and are free to list anything that they think was presented earlier. In such a task, participants might be presented with a photograph of a supermarket aisle containing an array of products. After viewing the photograph, followed by removal of it and a delay, they might be instructed to freely recall as many of the products as possible. With *cued recall*, the participant is specifically instructed as to what type of information to retrieve (e.g., Were you shown any cereal products? What is the brand of the box of cookies? What were the products in the bottom row?). In either case, the participant provides a listing (either specific words or brief summaries) of each item they can recall from the earlier presentation.

Memory protocols produced from recall tests can be analyzed in several ways, depending on what the researcher is hoping to learn from the content of the recalled material. The volume or quantity of memory (i.e., the number of different items listed) is sometimes used as a measure of the degree of attention and mental processing of the material presented earlier. A sparse listing of details suggests relatively little active processing of the stimulus materials, greater volume suggests more processing. For this purpose, it does not necessarily matter whether the “recalled” information was presented or reflects the perceiver’s own internal cognitions generated during the presentation stage (Petty & Cacioppo, 1986). For this reason, the sheer quantity of recall, without consideration of recall accuracy, is an imprecise measure of encoding that occurred during the original presentation, because we cannot know whether respondents are recording thoughts that they had at the time of the presentation, or thoughts that were generated later, during the recall task itself.

More often researchers are interested in the degree of *accurate* recall represented in the memory protocol, i.e., whether the items participants listed match information presented at the earlier time. For this purpose, the recall lists must be evaluated, and each item scored as correct or incorrect.³ The final measure may then be the number of items correctly recalled (overall accuracy of recall), or the researcher may be interested in which items were more likely to be accurately recalled (compared to information forgotten or incorrectly recalled). Research on this topic has shown that people tend to have superior recall for items near the beginning (*primacy effect*) and end (*recency effect*) of a list of previously presented items.

Finally, the researcher may be interested not only in the items correctly recalled, but also in the content of errors as represented by the items in the recall list that do not match information previously presented. Such incorrect items are referred to as memory “intrusions” and provide clues as to how the original material was encoded and interpreted before being stored in memory. Unfortunately, intrusion errors in recall protocols provide a very imperfect indicator of encoding processes. First, participants often are very cautious about what items they report on a recall measure (assuming that they are being tested for accuracy) and so do not list items unless they are fairly confident that they actually appeared. Consequently, the number of intrusions may be very small and unreliable as a measure of cognitive activity. Second, when intrusions do occur, we cannot tell whether they reflect cognitions that were generated at the time of the original presentation or simply bad “guesses” about information that cannot be recalled correctly.⁴ For this reason, recognition measures usually are more appropriate than recall measures for studying memory errors.

Another method of analysis of recall protocols involves the researcher noting the *sequencing* of the items recalled, specifically, which items are remembered first or later, and/or which items

are recalled together. The former provides information about accessibility in memory (i.e., which information is recalled most easily and rapidly, and which requires more search time and effort). The latter (measures of “clustering” in recall) provides information about how material has been *organized* in memory. Clustering measures are most often useful when information has been originally presented in some random or haphazard order but then appears in a different, more systematic order on the recall listings. Clustering measures are indices of the frequency with which items of the same “type” (or category) appear sequentially in the recall protocol compared to chance.⁵ These indices are used to document the kinds of categorizations perceivers use to encode and organize incoming information in memory. For example, Hamilton et al. (1980) used a measure of clustering in output of recall to show how behavioral information is organized in memory when perceivers are engaged in forming an impression about another person. In another social information processing study, Pryor and Ostrom (1981) used clustering measures to assess how incoming information about multiple persons in a social situation is processed and organized in memory. They found that when the persons were familiar individuals known to the perceiver, information was encoded and organized by individual person. But when the social stimuli were unfamiliar persons, memory was organized by behavioral categories rather than on a person-by-person basis.

Recognition measures. As a test of the content of memory, recognition does not require the respondent to retrieve items from the memory store, but to identify whether the information currently shown was among the materials presented on a prior occasion. The difference between recall and recognition measures of memory parallels the difference between an essay exam and a multiple-choice exam (Taylor & Fiske, 1981). As with a good multiple-choice exam, the researcher using recognition methods must carefully design and select options representing wrong answers (“foils”) that will appear to be correct if the respondent’s memory of the earlier material includes cognitions and assumptions that were not actually presented at the time. With recognition measures, the researcher’s interest is primarily in the types of errors that are made, rather than in the accuracy of memory.

In general, there are two different kinds of recognition tasks. In one paradigm, the respondent’s task is to review each item presented by the researcher and indicate whether that item of information was or was not seen before by responding “old” or “new” (or “true,” “false”). False recognitions (responding “old” to an item that was not present earlier) provide information about how the original materials were encoded and stored along with prior knowledge or inferences that the perceiver brought to bear at the time the information was received and processed. False recognition has been used to study the use of social category stereotypes in forming impressions of individual persons (e.g., Cantor & Mischel, 1977). The second type of recognition measure involves assessing memory confusions. In this case, participants are given two or more options and asked to indicate which one corresponds to information presented in the original materials.

An interesting application of the recognition confusion method is the who-said-what paradigm originally developed by Taylor et al. (1978). In this paradigm, participants first view an audiovisual presentation of a discussion among a group of six persons. The content of the discussion is presented on an audio recording while a picture of the individual group member who is speaking is projected on a screen. Later in the session, the participant is shown an array of photos of the group members and a series of sentences that occurred during the discussion. For each sentence, the participant is to choose which specific group member made that particular statement. What is of interest here is which group members are confused with each other when an incorrect choice is made. In the original experiment using this method, Taylor and colleagues varied the composition of the discussion group to determine whether discussants

were automatically categorized by sex or race while viewing the presentation. Consistent with the idea of automatic categorization, when the group consisted of three males and three females, recognition errors were more likely to involve confusing one male with another male or one female with another female (*intracategory* errors) than misattributing a statement made by a female to a male or vice versa (*intercategory* errors). Like clustering measures in recall, recognition confusions are used to assess how persons are classified or categorized in memory, even if they are unaware of this categorization process.

Both types of recognition task suffer from one major disadvantage as measures of what has been encoded and stored in memory. Both accurate and false recognition can reflect information that was encoded at the time of presentation, but they also can reflect guessing or inferences made by respondents at the time the memory measure is taken – that is, memory may be constructed (or reconstructed) when the person is tested on what he or she remembers. More sophisticated methods of analyzing recognition and recall errors make use of signal detection models to decompose memory scores and estimate the level of true recognition (Donaldson & Glathe, 1970).

Priming: Processing Without Awareness or Intent

Even though the memory measures discussed in the preceding section tap unintended memory, in the sense that participants are not aware at the time they receive information that they are going to be asked to remember it later, the memory tests themselves require active, conscious retrieval on the part of participants. Consequently, these methods can assess only what participants are aware of experiencing and willing to report to the experimenter. Other social cognition methods have been developed that do not require such an active role on the part of respondents to tap cognitive processing and other internal mental states. Specifically, *priming* techniques are designed to assess automatic cognitive and affective processes that occur without awareness or intent.

The concept of priming was introduced in Chapter 6 and refers to the unintended influence that recent or recurrent experiences have on subsequent thoughts, feelings, and behavior (Bargh & Chartrand, 2000). The idea underlying priming techniques is that exposure to a priming stimulus creates a state of mental readiness or preparedness for perceiving and interpreting subsequent information. Priming effects reflect implicit memory processes that function independently of what can be consciously retrieved from memory. Because these processes occur automatically and without awareness, the priming effect has come to be utilized to tap implicit cognition and affect.

Concept Priming

Priming studies usually involve two phases: (a) a priming task involving participant exposure to a stimulus (independent variable) (b) followed by an outcome judgment task (dependent variable) to assess the influence of the prime on subsequent ratings or judgments. Priming methodology was first introduced to experimental social psychology in a set of studies conducted by Higgins et al. (1977) that involved priming personality concepts and documenting effects on subsequent impression formation. The experimental procedures used by Higgins et al. (1977) are representative of the basic priming paradigm still used today. In the initial priming phase, participants performed a memorization task that exposed them to a set of words concerning either positive (e.g., “adventurous” and “independent”) or negative (e.g., “reckless” and “irresponsible”) personality traits. Then, in the subsequent judgment phase, the participants

took part in what they believed to be a separate, unrelated task during which they were asked to read a story about a person named Donald who was described as engaging in behaviors such as sailing across the ocean alone and preferring to study by himself. The description was written as to be ambiguous so that his behaviors could potentially be interpreted by participants in a relatively positive light (e.g., Donald is independent and adventurous) or negative way (e.g., Donald is reckless and irresponsible).

In the outcome phase, when participants were asked to report their impressions of Donald on a series of personality rating scales, those who previously had been exposed to words related to “adventurous” reported more positive impressions of Donald than did those who had been exposed to words related to “reckless” in the earlier task. This effect was obtained even though participants showed no awareness of the connection between the priming and judgment phases of the experiment.

Supraliminal priming. The research paradigm used in the Higgins et al. (1977) experiment involved a conscious priming task. During the first phase, participants were aware that they were viewing and processing words, even though they were unaware of the purposes of the word exposure. With such “supraliminal” priming, the participant is made fully aware of the priming stimuli but not of the underlying concept that the stimuli are intended to make accessible. In a study, suppose that you were presented with these words: “four,” “legs,” “sit,” “seat.” You are fully aware of these supraliminal words because they are presented explicitly on a page. The underlying emergent concept we sought to activate in your consciousness – but may not be completely obvious – is “chair.”

Another frequently used supraliminal priming task is the “scrambled sentence task” (e.g., Srull & Wyer, 1979). With this technique participants are given a cover story that the study is designed to measure their language ability and their task is to make coherent, grammatical sentences out of each of a string of words. For example: “him was about worried she always,” or “is dependent on occasionally he them” should be unscrambled as “she was always worried about him” and “he is occasionally dependent on them.” During the course of unscrambling the sentence priming task, the participant is exposed to words (e.g., “dependent,” “worried”) that are related to the concept that the researcher intends to activate (usually close synonyms). Later, the effect of the primed words on activating the concept is assessed on a subsequent impression formation or judgment task. Because the scrambled sentence task disguises the real purpose of the word presentation, participants are almost never aware that the earlier task has affected their subsequent judgment processes. Thus, supraliminal priming effects demonstrate how biases in person perception and decision-making can be invoked without participant awareness.

Subliminal priming. Awareness of the prime can be reduced even further by techniques that present the priming stimuli in a way that perceivers are not even conscious of seeing them. “Subliminal” exposure is achieved by presenting the prime (usually a word or a picture) very briefly for a fraction of a second, and then immediately masking this stimulus trace with a supraliminally presented neutral or nonsense stimulus. Subliminal priming was first used in a social psychology experiment by Bargh and Pietromonaco (1982) who used subliminal presentation to replicate the earlier trait concept priming experiments of Higgins et al. (1977) and Srull and Wyer (1979).

The key to subliminal priming is determining a time or duration of exposure of the priming stimulus that is too short to be consciously recognized. Usually, the stimulus is projected by a tachistoscope (a device developed by perception researchers to project stimuli at very brief exposures) or on a computer screen, with the participant gazing at a fixation point (e.g., an asterisk) at the center of the screen. The duration of the prime is a matter of milliseconds, although how long the exposure can be and remain below awareness depends on several factors, including

where in the visual field the stimulus is projected. With *foveal* processing, the priming stimulus is presented at the fixation point (within zero to two degrees of visual angle from the focal point of attention), a location at the center of the person's field of vision. With *parafoveal* processing, the prime is presented in the periphery or fringe of the visual field, at three to six degrees of visual angle from the focal point. Foveal presentation requires extremely short exposure time (on the order of 15 milliseconds) to be subliminal. Because parafoveal presentation is outside of the region of the focal point of attention, it allows for a somewhat longer duration (e.g., 60–120 milliseconds). However, it is somewhat more difficult to implement parafoveal presentations because the researcher must ensure that the participant's overall field of view includes the field in which the peripheral stimulus is presented.

Regardless of where the priming stimulus is presented, it must be followed immediately by a subsequent presentation (in the same location) of a “masking” stimulus to prevent extended exposure in visual iconic memory.⁶ A masking stimulus is a pattern with the same physical features as the prime. So, for example, if the priming stimulus is the word “PREGNANT,” the subsequent masking pattern would be a meaningless string of letters (“XQFBZRMQ”) that covers the location where the prime was presented. Participants who are subliminally exposed to this prime, intended to activate a sexist mindset, might be susceptible in the outcome judgment phase to answer “yes” to the question, “Are women less capable than men in management positions?”

Because subliminal exposure permits presenting primes without participants being consciously aware of the presented stimulus, there is no need to mislead them into believing that the stimulus exposure and judgment tasks are separate phases. Thus, the immediate effects of an activated concept on subsequent judgments or evaluations can be assessed. Subliminal priming has proved to be particularly useful for assessing the biasing effects of social stereotypes when perceivers are unaware that the stereotype has been subconsciously activated. Social category stereotypes have been primed by subliminal presentation of stereotype-related words (e.g., Devine, 1989; Devine et al., 2012) or photos of faces of out-group category members (e.g., Bargh et al., 1996, Experiment 3).

Assessing awareness. Regardless of whether supraliminal or subliminal priming techniques are used, it is important that the researcher determine that participants were truly unaware of the purpose of the priming manipulation. Awareness matters because if participants consciously recognize a connection between the presentation of the prime and the subsequent judgment task, they are likely to intentionally correct for the potential influences of the prime before making their responses in an attempt to appear unbiased. With supraliminal priming, the issue is whether participants become aware of the researcher's intent to activate certain constructs in the first task that may affect their judgments or behavior in the second task. To avoid awareness, it is important to camouflage the relation between the priming and judgment phases of the experiment as much as possible, including moving to different rooms, or having different experimenters give instructions.

Usually, manipulation checks for conscious awareness of the primes occur during an extensive debriefing session after the experiment is completed. Participants are probed for suspicions or knowledge of the intent of the experiment. This may be accomplished by using a “funneled debriefing” (e.g., see Chartrand & Bargh, 1996), a sequence of questions designed to elicit any suspicions or inferences that the participant may have made about the purpose of the experiment or the relationship between the priming task and the judgment task. When subliminal priming techniques are used, possible awareness of the prime presentations can be assessed by somewhat more objective means. The recommended procedure is to tell participants at the end of the experiment that they were exposed subliminally to some stimuli and ask them to try to guess the content of the presented stimuli. If they are unable to explicitly recall any of the words or

images that were presented as primes, it is safe to assume that they were not consciously aware of the content of the material used in the priming exposure. An even more conservative procedure is to give participants a multiple-choice recognition test to see if they can identify the actual primes from among a set of distractor foils. (Because participants may correctly guess some of the correct answers by chance, it is best to compare their performance to that of a control group of respondents who have not actually been exposed to the subliminal stimuli. If the participants exposed the priming stimuli are unable to guess more accurately than the controls, it is safe to assume that awareness was not a problem.)

Sequential Priming

Another variation on priming techniques is used to assess automatic associations between mental concepts. The idea behind sequential priming is that if one stimulus has become associated with some other concept, feeling, or behavior, then presentation of that stimulus will automatically activate (prime) those associations. In that case, if the prime and the association are presented sequentially, responding to the second (target) stimulus will be facilitated because of the prior preparation produced by the prime.

The basic structure of the sequential priming paradigm is as follows. On each trial, the prime stimulus (a word or a picture) is presented for a short duration (e.g., 150 milliseconds), then erased, and after a brief delay, the target stimulus (outcome task) is presented, and the participant makes a judgment about the target by pressing a key to indicate his/her response. The outcome measure of interest is the speed of reaction time to make a judgment of the target stimulus. If the target is associated with the prime in a person's memory, then responding to that target should be facilitated when the prime has been presented just before. Thus, if the response made by a participant is faster when the target (e.g., tangerine) is preceded by a relevant prime (e.g., orange) than when the same target is judged in the presence of an irrelevant prime (e.g., carpet), this indicates that the two concepts (orange and tangerine) are automatically associated in a person's memory. Because reaction times are measured in milliseconds, sequential priming requires the use of software programs with very precise timing recorders to detect minor (millisecond) differences in average speed of response.

Although the duration of presentation of the priming stimulus in the sequential priming paradigm is relatively short, it is a supraliminal exposure and hence participants are consciously aware that they perceive it. For this reason, sequential priming requires some cover story that explains to participants why they are viewing the primes even though their task is to respond only to the target stimuli. This may be done by telling participants that the study is about people's ability to perform two tasks simultaneously. On each trial, they are being asked to attend to and remember the first stimulus presented while at the same time they are to make a judgment about the second (target) stimulus. In addition, the potential effects of awareness are minimized by limiting the amount of time for making a response. The window of time allowed between the onset of the prime presentation and the onset of the subsequent target presentation (the "stimulus onset asynchrony," or SOA) is kept short (usually no more than 300 milliseconds), so that there is no opportunity for conscious processing of the priming stimulus before a judgment about the target is required. With such brief delays, only automatic (implicit) effects should be able to occur.

Lexical decision task. In one version of the sequential priming technique, the target stimulus is a string of letters and the judgment that the participant is to make is to indicate as quickly as possible whether the string is an actual word or not. If the target is a word, this judgment is made more quickly when it is preceded by presentation of a prime concerning a related word or

concept. For instance, when the prime is an overarching category label (e.g., “furniture”), target words representing members of that category (e.g., “chair”) are recognized faster than target words that do not belong to the category (e.g., “bird”) (Neely, 1977). Based on this principle, the lexical decision task can be used to assess automatic activation of stereotypic traits when the prime is a social category label or picture (e.g., Cunningham & Macrae, 2011; Macrae et al., 1995).

Automatic evaluation. The sequential priming technique was adapted for use in social psychology by Fazio and his colleagues to assess automatic attitudes or evaluative responses to social stimuli (Fazio et al., 1986). In Fazio’s version of the paradigm, the priming stimuli are words or pictures of an attitude object, followed by target words that are evaluative adjectives (e.g., “delightful,” “awful”). The respondent’s task is to indicate as quickly as possible whether the target word has a good or bad connotation. If the prime automatically elicits an evaluative reaction of some kind, then the participant’s evaluative response is expected to carry over to the judgment of the target. If the evaluative meaning of the target sufficiently matches that of the prime, responding should be facilitated. So, for instance, if the primed stimulus has a positive connotation (e.g., “party”), it should then speed up judgments of positive target adjectives (e.g., “good”). Conversely, a positive prime should slow down (inhibit) judgments of a subsequent negative target adjective (e.g., “bad”). Thus, the pattern of facilitation and inhibition of evaluative judgments of a target provides an indirect (implicit) measure of attitudes activated by the prime. Presentation of a positive prime will speed up responding to positive adjectives and slow down responding to negative ones. Presentation of a negative prime will speed up responses to subsequent negative judgments and inhibit positive judgments.

Fazio’s automatic evaluation paradigm has been used specifically to measure implicit prejudices. For example, Fazio et al. (1995) used photos of White and Black faces as primes in the sequential priming task. Participants were instructed to attend to and remember the faces (for a later recognition test) while they were doing an adjective evaluation task (target stimuli). On some trials, positive or negative adjectives were preceded by a prime of a facial photo of a White person; on other trials the same adjectives were preceded by a prime of a photo of a Black person. Implicit prejudice is assessed by comparing (for each participant) the response times on trials with Black primes to those obtained on trials with White primes. Automatic negative attitudes toward Blacks are indicated when judgments of negative adjectives are faster following Black primes than following White primes. Conversely, automatic positive attitudes toward Whites are indicated when judgments of positive adjectives are slower for Black primes than White primes. Interestingly, individual differences in automatic evaluation were not correlated with explicit self-report measures of racial prejudice, suggesting that the priming task taps into evaluative reactions that people may be unwilling or unable to report directly.

Issues related to use of reaction time measures. All the sequential priming techniques (as well as several other social cognition methods) rely on the analysis of reaction times as an indicator of automatic processes. Researchers making use of these methods need to be aware of some methodological issues in using reaction times as a dependent variable. Many factors other than the priming effects of interest can influence response latencies to target stimuli, including word length and word frequency. Thus, it is extremely important that these stimulus features be controlled for in making comparisons between priming conditions.

Reaction time measures also create some problems for data analysis purposes. First, the distribution of response times is typically positively skewed (in that very long reaction times occur occasionally but extremely short latencies are impossible). For this reason, various transformations of the reaction time measure (e.g., taking the square root, or the natural logarithm) may need to be used to normalize the distribution for purposes of analysis. Second, the researcher

needs to be concerned about “outliers” in each participant’s reaction time data; excessively long reaction times indicate the respondent wasn’t paying attention at the time of presentation of the target stimulus and excessively short reaction times reflect anticipatory responding before the target was fully processed. Such outliers should be removed from the data set before analyses are conducted. Typically, response times shorter than 300 milliseconds are trimmed as too fast to represent true responses (Bargh & Chartrand, 2000). In general, only very extreme outliers should be trimmed (e.g., reaction times that are more than three standard deviations above and below the mean reaction time), and care should be taken that deletions are equally distributed across the different priming conditions.

Other Measures of Automaticity

Sequential priming methods are designed to assess responses elicited spontaneously or automatically, without intent or effort on the part of the respondent. Most automatic responding is presumed to occur early in information processing; given additional time and cognitive effort, some automatic processes may be overridden or corrected by more thoughtful, deliberative cognitive processing (Conrey et al., 2005; Payne, 2001). Hence the importance of short SOA in the sequential priming paradigm, as brief response times preclude the opportunity for deliberative processing. Alternative methods for assessing automatic responses rely on different strategies for reducing the influence of intentional processing.

Cognitive busyness. Capacity for conscious processing can be limited by various techniques for creating cognitive overload, either presenting a lot of information very rapidly (e.g., Bargh & Thein, 1985), or occupying cognitive resources with a secondary task (what Gilbert et al. (1988) referred to as “cognitive busyness”). A cognitive busyness manipulation might require participants to hold an eight-digit number in memory (distractor task) as they are engaging in a judgment task or while the stimulus information is being presented (the primary task: see Parker et al., 2012). Gilbert and Osborne (1989) used this method to test a two-stage model of attribution processes. Participants in the experimental condition were given a string of numbers to remember throughout the entire time they were viewing a videotape of a woman engaging in a discussion with a stranger; participants in the control condition viewed the same tape without the secondary memory load task. After the video presentation, all were asked to make attributions about the causes of the actor’s behavior. As predicted, participants in the cognitive overload condition gave more dispositional (e.g., personality) attributions than those in the control condition, who provided more situational explanations for the behavior. These findings were consistent with the theory that dispositional judgments occur automatically and effortlessly, whereas situational attributions require more deliberation and cognitive effort. In later studies, Gilbert and Hixon (1991) used a similar cognitive overload manipulation to assess automatic activation and use of social stereotypes in impression formation.

Response interference techniques. Responses that are automatically elicited can be expected to interfere with production of other responses that are incompatible with it. We have already mentioned the use of interference effects in connection with the Stroop color-naming task as a measure of automatic attention allocation. The Stroop effect demonstrates that performance of an instructed response can be inhibited when an incompatible automatic response is elicited by the stimulus. The occurrence of such interference can be interpreted as an indication of automatic processing at work.

As another use of response interference as a measure of automatic processes, the **implicit association test** assesses automatic associations between mental concepts by classifying and sorting items as quickly and accurately as possible into different categories (Greenwald et al.,

1998). Administered using a computer, a stimulus (word, person's name, or picture) is presented in the center of the screen, and the respondent is instructed to press a key to indicate whether it is an exemplar of one of two broader categories. The experiment is designed so that a phase involves making judgments using response pairings that are stereotypically congruent, and another phase in which response pairings are stereotype incongruent. These two phases are usually counterbalanced across participants.

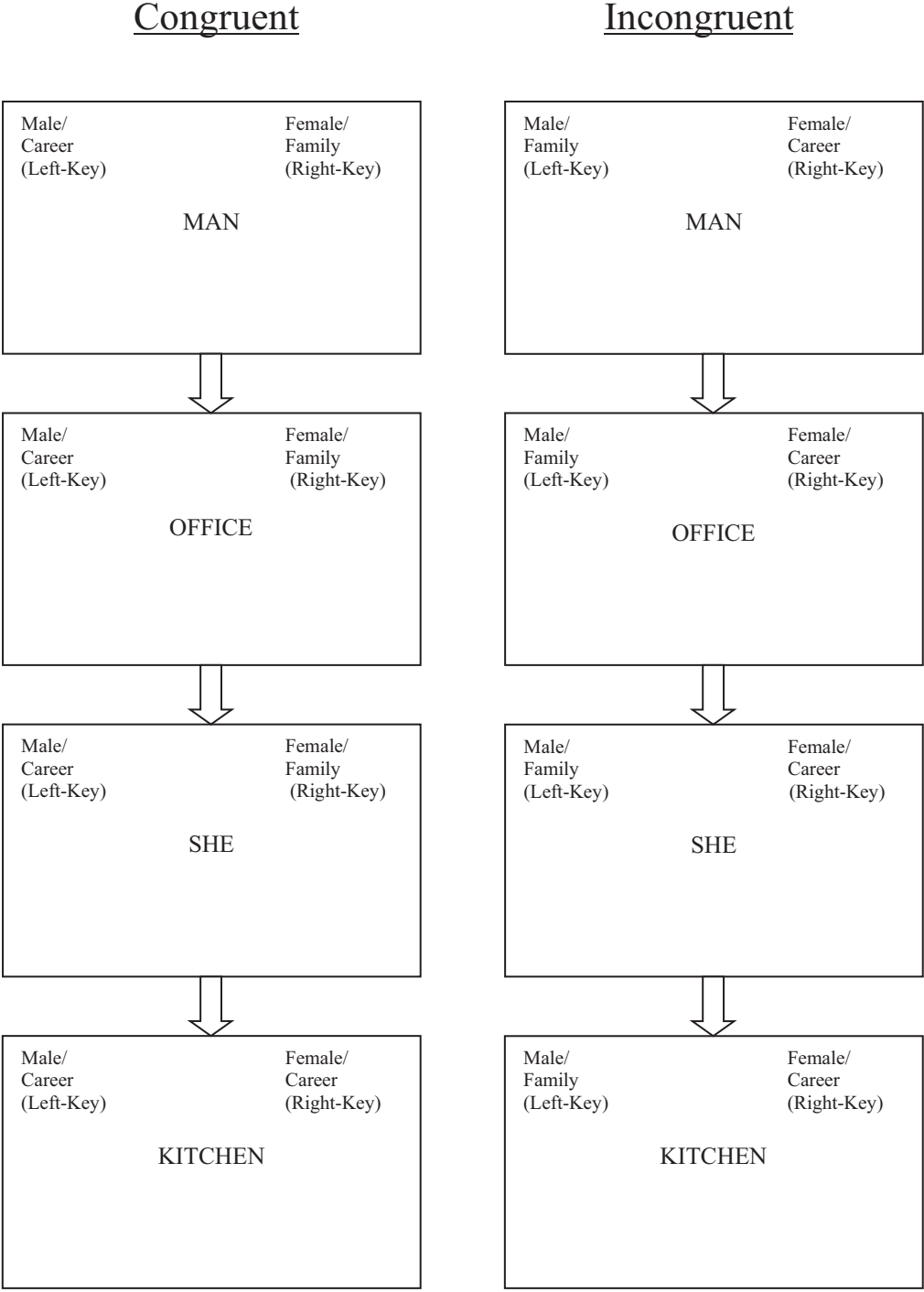
An example of an IAT, with four trials in each of the congruent and incongruent conditions, is presented in Figure 18.2. In the congruent phase, the participant is instructed to press the right-hand key if the word shown at the center is either Female or Family-related, and the left-hand key if it is either Male or Career-related. In the incongruent phase, the pairings are changed so that the right-hand key should be used if the word shown is either Female or Career-related and the left-hand key is used for either Male or a Family-related word. If gender-stereotypic roles are mentally associated in a person's memory, then trials where the response pairings are stereotype congruent (e.g., Male/Career) will have faster reaction times than trials in which the pairings are reversed and stereotype incongruent (e.g., Male/Family). That is, people who hold implicit gender stereotypes are expected to respond more quickly to trials consisting of response pairings that are consistent rather than inconsistent with these stereotypes. Thus, the difference in speed of responding of the congruent over the incongruent trials is a measure of degree of implicit association between concepts represented in a person's memory. The IAT has been widely used in social psychological research to assess implicit prejudice (associations between social categories and evaluative words) and stereotype content (associations between social categories and stereotypic trait words).⁷

A quantitative review examined 126 previous studies that have assessed the degree of correspondence between IAT measures and the equivalent versions of self-report measures (Hoffmann et al., 2005). The average study reported a correlation of .24, suggesting that although explicit and implicit measures tended to coincide, other underlying mechanisms are responsible for making these measurement methods largely distinct. More recent research has raised the issue of the validity of the IAT; the predictive validity of this methodology is not without its detractors (Oswald et al., 2013; Schimmack, 2021).

Social Psychophysiology: Physiological Traces of Affect and Cognitive Processing

The various priming techniques described in the preceding section have the advantage of tapping processes and subtle responses that may not be accessible to conscious awareness. However, they all share the disadvantage that the researcher must be concerned that participants not become aware of the true purpose of the priming manipulation, because awareness can influence the critical responses. Internal physiological responses are less susceptible to alteration by conscious awareness and thus provide the possibility for yet less reactive methods for assessing implicit, unintended responses.

The use of measures of physiological responses has a long history in social research, as seen in research designed to assess internal states such as interpersonal attraction, aversion, specific emotions, or stress. Most of these early investigations relied on unitary physiological measures such as respiration, pulse rate, finger temperature, and skin conductance (see Blascovich et al., 2001; Guglielmi, 1999). Researchers were relatively unsophisticated about underlying physiological processes and as a consequence most of these early attempts did not result in valid or reliable assessments of the psychological states of interest. One major problem is that there is rarely, if ever, a simple one-to-one relationship between a single physiological response and some specific



internal psychological state (Blascovich, 2000; Blascovich & McCall, 2013). The often-used Galvanic skin response (GSR), for instance, provides an indicator of general arousal, which may arise from any number of positive or negative states including stress, excitement, aversion, or sexual desire. Changes in heart rate may signal physical exertion, attention, anticipation, or arousal, with relatively little differentiation when measured in isolation of other physiological responses.

Fortunately, recent advances in theory and methodology now allow for multivariate physiological assessments that are capable of distinguishing different motivational states, positive and negative affect, attention, and active cognitive processing. Because these indicators are potentially relevant to many phenomena of interest to social researchers, such as assessment of prejudice, social emotions, social stress, and interpersonal motives, the availability of sophisticated physiological measures has given rise to the subfield of “social psychophysiology” (Cacioppo & Petty, 1983) or “social neuroscience” (Cacioppo et al., 2000; Norman et al., 2013) designed to capitalize on these methods for understanding the interrelationships between psychological and physiological/neurological states. The downside of this development is that these techniques require expensive equipment, sophisticated programming, and more than superficial training in physiology, neuroscience, and complex statistical techniques to make good use of the available methods. Nonetheless, the capability for conducting experiments that include physiological measures is gradually expanding, and physiological methods are taking their place among our tools for assessing implicit cognition and feelings.

Most physiological measures used in social research involve noninvasive techniques, that is, ways of recording internal physiological responses from the surface of the body. Usually, some form of electrophysiological recording is involved because various physiological events produce detectable electrical signals at the surface level (Blascovich, 2000). Depending on the location and type of recording, electrophysiological measures record signals of neural, cardiovascular, or endocrine system activity occurring in response to current situational inputs. In modern methods, it is the patterns of responses from multiple signals over time that serve as markers of specific psychophysiological states. Some examples of uses of these physiological measures in social research will be described in the following sections.

Cardiovascular Indices

Although heart rate alone is an inadequate marker of any specific arousal state, measures of heart rate in combination with other recordings of cardiac and vascular performance can differentiate different states of anticipation or stress (Blascovich & Kelsey, 1990). Tomaka et al. (1993), for example, demonstrated that specific patterns of cardiovascular responses can distinguish between feelings of threat (a negative, avoidance state) versus challenge (a positive, approach state) as motivational states in anticipation of potentially difficult or stressful situations such as performing difficult arithmetic problems or preparing for a public speech. Threat is marked by increased pituitary-adrenocortical activity (PAC), which is associated with negative physiological consequences of stress. A challenge response to the same situation, however, is marked by increased sympathetic-adrenomedullary activity (SAM), which is associated with benign states and improved performance (Dienstbier, 1989).

Tomaka and Blascovich (1994) validated the usefulness of this distinction for assessment of psychosocial processes by demonstrating that individuals with high beliefs in a just world (as measured by self-report questionnaire) showed a challenge response to a potentially stressful performance situation and performed better than individuals with low beliefs in a just world, who instead showed a threat response pattern to the same situation. In another study, Blascovich

et al. (2001) found that White participants showed a predominance of threat response when assigned to work in a team task with a Black partner, compared to participants who showed a challenge response when working with a same-race partner.

Facial Measures of Affective States

One of the difficulties of many physiological indicators of arousal is that these measures do not distinguish between arousal due to positive affect or approach and negative affect or avoidance. Facial expressions, however, do vary in ways that correspond to specific positive or negative social emotions (Ekman, 1993). Ekman and Friesen (1978) developed and validated an elaborate facial action coding system (FACS) to assess various emotional states based on detailed aspects of spontaneous facial expressions. Using this system, researchers can be trained to observe people's facial expressions to distinguish among the basic emotions of happiness, sadness, anger, fear, disgust, contempt, and surprise, which Ekman (2007) considered universal across cultures. This method of measuring internal states has several drawbacks, however. First, training to use the FACS is extensive and requires a considerable investment of time and effort. Second, although facial expressions are usually produced without conscious intention, it is possible to control and conceal facial responses if a participant makes an effort to do so. Thus, facial expression is not always a valid measure of implicit affect.

Facial electromyograph. Although overt facial expressions are potentially controllable, individuals are not able to control the tiny, visually imperceptible movements of specific facial muscles that occur at the onset of an emotional or affective response to stimulus events. Thus, Cacioppo and his colleagues (e.g., Cacioppo & Petty, 1981; Cacioppo et al., 1986) have recommended the use of facial electromyograms (EMG) specific to targeted facial muscles as physiological markers of positive and negative affective states. EMG measures focus on specific muscles of the eyes and mouth associated with corrugator ("frown muscles") and zygomaticus ("smile muscles") activity. Electromyographic recordings obtained from electrodes distributed across areas of the face indicate that corrugator EMG activity increases and zygomaticus activity decreases during negative affect. Conversely, zygomaticus EMG increases and corrugator EMG decreases signal positive affect (Cacioppo et al., 1986). Interestingly, the EMG recordings successfully discriminate between different affective states even when judges observing videotapes of the participants' overt facial expressions are unable to identify whether the affect is positive or negative.

Vanman et al. (1997) made use of the diagnostic capability of EMG activity to assess implicit attitudes toward members of racial outgroups. In their experiments, White participants were shown slides of Black or White individuals and were asked to imagine themselves in a situation in which they were partnered with that individual in a cooperative task. On overt self-report ratings of the potential partners, respondents tended to show preferential ratings for Black targets. EMG facial measures, however, showed the opposite pattern, with more negativity exhibited toward Black partners.

Startle eyeblink reflex. Another minute facial muscle measure that may prove useful to index affective states is electromyograms specific to reflexive eyeblinks (Blascovich, 2000). The startle eyeblink response refers to the reflexive blinks that occur when individuals perceive an unexpected, relatively intense stimulus, such as a loud sound. The startle eyeblink reflex is negatively toned. Hence, Lang and his colleagues (1990, 1992) have reasoned that the eyeblink response should be facilitated or enhanced if the perceiver is in a negative affective state and inhibited if the perceiver is experiencing ongoing positive affect. Subsequent research has demonstrated that the strength or intensity of the startle eyeblink reflex as measured by EMG activity

is sensitive to affective states as predicted. Thus, the eyeblink reflex may prove useful, as well as other EMG indices, as a measure of covert affect.

EMG techniques require sophisticated equipment and technology to implement and have the added disadvantage that participants must sit still (with electrodes planted on various parts of their faces) and minimize extraneous head movements while responses are being recorded. Thus, the method is hardly unobtrusive and the kinds of stimulus situations in which the technology can be used are limited to relatively passive viewing conditions such as presentations of slides or video displays. Nonetheless, EMG recordings have proved to be valid indicators of implicit affect that individuals may be unaware of – or unwilling to express overtly. EMG measures can be a useful tool for testing theories about activation and influences of nonconscious affect, even if their practical application may be limited.

Measures of Brain Activity

Indicators of cognitive and affective processing are also assessed in various noninvasive measures of brain activity. Again, these techniques require access to expensive and sophisticated equipment (much of it developed for medical research), but they are finding their way into social research applications. Neuroscientific or brain imaging procedures to assess brain activity are classified into three major families: transcranial magnetic stimulation, event-related potential, and functional magnetic resonance imaging (Quadflieg & McCrae, 2011). A summary description of the purpose of each neuroscience method, its system of measurement, and its assumptions are presented in Table 18.1. Although each neuroscience technique is tailored to address a general type of research question, it is possible to employ all three to study the same brain processes.

Brain stimulation. In **transcranial magnetic stimulation** (TMS), external magnetic pulses are transmitted through the scalp to different brain regions, to identify the part of the brain that produces a specific behavior. By sending magnetic pulses, neurons in these regions are

TABLE 18.1 Major neuroscientific methods

<i>Technique</i>	<i>Purpose</i>	<i>Measurement</i>	<i>Assumption</i>
Transcranial magnetic stimulation (TMS)	Examine cause and effect between stimulation of a brain region and the consequent reaction	Measures behavioral reaction due to magnetic field stimulation of a brain region. Data are considered experimental.	Brain processing invoked artificially by a magnetic field is the same as what naturally occurs.
Event-related potential (ERP)	Examine rapid changes in brain wave activity during mental operations	Measures electrical activity in brain using electrodes placed on the scalp. Amplitudes and shapes of different varieties of brain waves may be examined. Data are considered correlational.	Brain processing is represented by greater electrical activity in brain.
Functional magnetic resonance imaging (fMRI)	Examine and map brain regions active during mental operations	Measures level of oxygen flow in the blood vessels of the brain. Data are considered correlational.	Brain processing is represented by greater oxygen flow in blood vessels.

excited from their resting state. Because a brain locale is stimulated with focused magnetic fields, the technique is primarily undertaken to determine the cause-and-effect relation between brain regions and behavioral responses. In one TMS study, Cattaneo et al. (2011) examined gender stereotyping. Two areas of the frontal cortex – the left dorsolateral and right anterior – had been implicated in prior research as related to stereotypical reactions. Magnetic waves were targeted to both these areas while participants performed an implicit association test regarding gender stereotypes. When these regions were stimulated, compared to when they were not, male participants exhibited greater anti-female biases on the IAT. Female participants did not exhibit a gender bias on the test. Generally, in TMS research, the technique is coupled with the use of event-related potentials or functional magnetic resonance imaging, to verify that the targeted brain region was indeed stimulated after invoking the magnetic wave.

Brain wave measures. In research using techniques of **event-related potential (ERP)**, electrodes attached to multiple regions of the scalp are used to measure rapid waves of electrical brain activity arising from the processing of information. An electrical storm of voltage fluctuations is generated while a participant is processing information. Examination of the amplitude of ERP is typically assessed with the procedure known as electroencephalography (EEG), with these terms used interchangeably in the literature. In typical ERP studies, participants are presented with a series of stimuli to process, and then various types of brainwave signals are gauged.

Use of ERP measures in social research include studies of hemispheric asymmetry (differential activation of the right or left hemispheres of the brain under different stimulus conditions or tasks) and computation of the patterns of EEG waveform that follow presentation of a stimulus. In both types of measures, the location and timing interval of differential brain activity is used to assess emotional responding, categorization, and evaluative reactions (Guglielmi, 1999). For example, Cacioppo and his colleagues (e.g., Cacioppo et al., 1993; Cacioppo et al., 1994) have made use of one component of ERP – the late positive potential (LPP) which occurs approximately 300 milliseconds after onset of a stimulus – to assess implicit positive and negative evaluative reactions. The amplitude of the LPP is larger when a presented stimulus is unexpected or categorically distinct from previously presented stimuli. Thus, to measure whether a stimulus is reacted to negatively, participants are first presented with a series of (known) positive stimuli and then the target stimulus. If ERP amplitude increases significantly in response to the target, it is assumed that a shift to negative evaluation has been registered, even when respondents are unwilling (or unable) to report a negative evaluation overtly.

Neuroimaging. Advances in brain imaging techniques have seen widespread application in the field of cognitive neuroscience and are gradually finding application in social research as well. Regional cerebral blood flow, as measured by *positron emission tomography* (PET), has been used to investigate the relationship between regional brain activity and induction of various emotions, but different studies have produced discrepant findings about the regional differentiation of specific emotions (e.g., Lane et al., 1997; Robinson, 1995). PET scans also have been used to assess differences in processing of information relevant to the self compared to processing of information about others (Craik et al., 1999).

More recently, **functional magnetic resonance imaging** (fMRI), which measures the relative degree of oxygen flow through blood vessels in the brain, has been used to identify which brain regions are implicated in various cognitive and emotional processing functions. Fluctuations in low and high oxygen flow throughout the brain, as a result of information processing, are detected by placing a participant in a magnetic resonance scanner. Software programs assist researchers by showing the degree of oxygenation in each brain subregion. Unlike a PET scan, the methods of fMRI do not require participants to ingest some sort of radioactive biochemical substance to trace the movement of a chemical in the brain during mental

processing. This is a main reason why fMRI has virtually supplanted PET scanning research. An assumption of fMRI is that a one-to-one correspondence exists between greater brain activity and greater oxygen flow in a locale. Although this correspondence is strong, as cerebral blood flow is indeed linked to brain oxygenation in those same areas, it is possible that processing of some types of information might not attain a sufficient threshold to be biologically detected and indicated in the form of oxygen flow.

Interpreting the results of fMRI output for social research requires a deep understanding of what activation of specific regions of the brain signifies. For example, one intriguing finding with potential social implications is fMRI evidence that processing faces of racial out-group members generates more activity in the amygdala region of the brain than processing faces of in-group members. Further, White individuals whose amygdalas fired up the most also scored higher on two other measures of implicit attitudes toward Blacks (Phelps et al., 2000). This result was interpreted in light of other evidence that the amygdala is an area of the brain implicated in negative emotional reactions such as fear and anxiety. However, more recent research indicates that amygdala activation has many different processing implications and, more generally, signals the motivational importance of social stimuli (Van Bavel & Cunningham, 2010). Thus, although exploring the brain activity associated with social processing is generating a great deal of interest, we caution against expecting any simple one-to-one relationships between specific brain regions and complex cognitive and evaluative processes (Willingham & Dunn, 2003).

The output of all three families of brain imaging studies is subject to considerable statistical variation across measurement occasions within the same individual. A neutral stimulus (e.g., blank screen) is often used as a base assessment of random brain activity. This is necessary because the brain remains active even when a participant is not actively processing information or is distracted by random thoughts. Measurable difference in brain activity beyond that of the neutral stimulus may be attributed to processing of the stimulus (signal). A large signal relative to baseline indicates brain reaction and processing of the presented stimulus. Another strategy to rule out interference due to random brain noise is to repeatedly present the same stimulus in sequential trials to the same participant. The patterns of brain activity are then averaged across the multiple trials. The idea is that random noise should cancel out after averaging the many recordings of brain reactions. Brain imaging studies commonly use repeated-measures designs, with each participant serving in the control condition (neutral stimulus) as well as in the treatment condition (target stimulus).

Statistical variations in brain activity also may fluctuate across people. Because participants arrive at the research laboratory with different preexisting memories and histories, even if viewing the same stimulus, different people might produce slightly different patterns of neuroactivation. Adding to this complexity is that more than one region of the brain could be activated when perceiving a stimulus. However, by conducting social neuroscience research using many participants who are presented with the same stimulus, it is possible to determine an “average” or typical pattern of neuroactivation.

Interpreting imaging patterns is still in early stages of development. A frequent question raised about neuroimaging research is the extent to which such studies have external validity or generalizability outside the laboratory (Quadflieg & Macrae, 2011). Given that this research modality typically requires participants to lie down on their backs in a still pose in a confined space while neuroimaging instruments are recording brain activity, subtle head movements will inaccurately isolate the cortical region activated upon receipt of a stimulus. Obviously in natural everyday situations, when processing social information, people usually are not physically or socially restricted in such unusual ways. Furthermore, the types of stimulus and response tasks used in neuroscience methods are relatively artificial and systematically controlled through

computer presentation. To illustrate the unusual social context of a brain imaging study, a presented stimulus might involve a headshot of a person of the opposite sex. This unidirectional process of the participant observing and having a brain response to a static photograph of a head does not mirror the reality of interpersonal social exchanges, which typically include the brain functions of listening, speaking, and physical gesturing. Current technology primarily permits imaging instruments, which are computer intensive and bulky, to be conducted inside the confines of a laboratory. Thus, laboratory research using neuroimaging measurements is susceptible to many of the same advantages (e.g., systematic control of variables) and disadvantages (e.g., artificiality) of laboratory studies in general.

Conclusion

The social cognition methods reviewed in this chapter are intended primarily to assess implicit cognitions and feelings that respondents are unwilling or unable to report on more explicit measures of internal states. Measures of attention, processing time, memory, reaction time, and physiological responses provide a different perspective on mental processes than that provided by traditional self-report methods. When these different types of assessment all converge on the same diagnosis (as evidenced by strong correlations), we have impressive convergent validation of the construct under investigation. But as we have seen, implicit and explicit measures of internal states sometimes produce discrepant findings. Such differences in outcome raise the issue of which methods assess the “true” attitude, feeling, or cognitive process most accurately.

Possibly, discrepancies between explicit and implicit measures of attitudes, feelings, and beliefs reflect the fact that respondents are hiding or misrepresenting their true responses on overt self-report measures. In that case, implicit measures provide a more valid assessment of participants’ actual thoughts and feelings, because these are less susceptible to intentional control or deception. And, indeed, it often is the case that discrepancies between implicit and explicit measures occur with respect to assessments of attitudes or beliefs that are undesirable or politically sensitive, such as racial prejudices, stereotyping, or negative emotional states. Another possibility for differences between responses on explicit and implicit responses is not because people deliberately misrepresent their conscious beliefs or feelings, but because their conscious attitudes are in actuality different from their subconscious responses. Explicit attitudes that derive from controlled cognitive processing may reflect what individuals consciously and intentionally believe and value, even though they may hold residues of more negative affective reactions and beliefs at the level of automatic, unintentional processes. In this case, explicit self-report measures may be more valid than implicit measures, if one presumes that most of social life is carried out under consciously controlled processing conditions.

Most researchers who use implicit measures do not contend that such measures are more valid than traditional self-report measures of respondents’ true mental or affective states. Rather, most believe that implicit or automatic responses reflect underlying processes different from those assessed via explicit measures. The challenge is to determine what processes and outcomes are related to these implicit measures that are not predicted or accounted for by other methods of assessment. As we have discussed, automatic processes may be more likely to emerge when individuals are performing under cognitive overload or extreme time pressure. There is also evidence that implicit measures of automatic reactions predict subtle nonverbal behaviors in certain social situations better than do explicit measures. For example, in the study we described previously by Fazio et al. (1995) that used sequential priming to assess automatic racial prejudice, the reaction time assessment of automatic evaluation was not correlated with participants’ scores on a self-report measure of racial attitudes. However, when participants interacted with a Black

confederate of the experimenter after the attitude measurement session, subsequent ratings by the confederate of the participant's friendliness and interest during the interaction proved to be significantly correlated with the individual's evaluative bias in the priming task. Thus, the automatic evaluation measure did seem to successfully predict other subtle, nonverbal (and probably nonconscious) overt behavior. Other studies of the predictive validity of implicit measures such as this will be needed to better understand the relationship between implicit cognitive and affective processes and social behavior.

Notes

- 1 Actually, eye fixations shift three or four times a second within minute distances even while attending to a single object. However, this level of precision is not appropriate for most social research purposes, where more gross assessments of direction of gaze are sufficient.
- 2 The time interval between presentation and recall can vary from just a few minutes to a matter of hours or days, depending on whether short-term or long-term memories are being assessed. When the recall measure is taken in the same session within minutes of the original presentation, participants are usually given some unrelated "filler" task to occupy their attention during the interval and prevent active rehearsal of the presented materials. When the recall measure occurs at some later time in a separate session, the researcher has less control over the intervening events and cannot know whether the participant has been thinking about or rehearsing the presentation during the lapsed period of time. However, because participants had not been instructed to remember what they experienced, it is generally assumed that such rehearsal is unlikely.
- 3 Different standards of accuracy may be applied depending on the type of material being recalled. When very specific items of information have been presented, an exact match may be required for correctness. However, when the information is more complex or ambiguous, items are often evaluated by the "gist" criterion. That is, an item is scored as correct if the coder judges that it captures the general idea of what was presented.
- 4 The same problem applies to interpretation of accurate responses, because we cannot know whether correct items represent actual memory of the presented materials or simply good "guesses." However, the more detailed and complex the information that has been presented and recalled, the less likely it is that guessing accounts for accuracy.
- 5 The two most commonly used clustering measures are the Stimulus Category Repetition index (SCR) developed by Bousfield and Bousfield (1966), and the Adjusted Ratio of Clustering (ARC) measure recommended by Roenker et al. (1971). See Hamilton et al. (1980) and Ostrom et al. (1980) for discussions of the relative merits of each of these indices for social cognition research.
- 6 Unless the visual buffer is erased or overwritten, a visual image remains in short-term memory store even after the stimulus image has been removed from the visual field.
- 7 If the reader is interested in how the IAT is conducted, IAT tests of implicit associations to race, sexual orientation, gender, and age can be accessed on the Internet at <https://implicit.harvard.edu/implicit/>.

References

- Atalay, A. S., Bodur, H. O., & Rasolofoarison, D. (2012). Shining in the center: Central gaze cascade effect on product choice. *Journal of Consumer Research*, 39, 848–866.
- Bargh, J. A., & Chartrand, T. L. (2000). Studying the mind in the middle: A practical guide to priming and automaticity research. In H. Reis & C. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 253–285). New York: Cambridge University Press.
- Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, 71, 230–244.
- Bargh, J. A., & Pietromonaco, P. (1982). Automatic information processing and social perception: The influence of trait information presented outside of conscious awareness on impression formation. *Journal of Personality and Social Psychology*, 43, 437–449.
- Bargh, J. A., & Thein, R. D. (1985). Individual construct accessibility, person memory, and the recall-judgment link: The case of information overload. *Journal of Personality and Social Psychology*, 49, 1129–1146.

- Blascovich, J. (2000). Using physiological indexes of psychological processes in social psychology. In H. Reis & C. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 117–137). New York: Cambridge University Press.
- Blascovich, J., & Kelsey, R. M. (1990). Using cardiovascular and electrodermal measures of arousal in social psychological research. In C. Hendrick & M. Clark (Eds.), *Research methods in personality and social psychology* (pp. 45–73). Newbury Park, CA: Sage.
- Blascovich, J., & McCall, C. (2013). Social influence in virtual environments. In K. E. Dill (Ed.), *The Oxford handbook of media psychology* (pp. 305–315). New York: Oxford University Press.
- Blascovich, J., Mendes, W., Hunter, S., Lickel, B., & Kowai-Bell, N. (2001). Perceiver threat in social interactions with stigmatized others. *Journal of Personality and Social Psychology*, 80, 253–267.
- Bousfield, A. K., & Bousfield, W. A. (1966). Measurement of clustering and sequential constancies in repeated free recall. *Psychological Reports*, 19, 935–942.
- Brewer, M. B., Dull, V., & Lui, L. (1981). Perceptions of the elderly: Stereotypes as prototypes. *Journal of Personality and Social Psychology*, 41, 656–670.
- Cacioppo, J. T., Berntson, G. G., Ernst, J. M., & Ito, T. A. (2000). Social neuroscience. In A. E. Kazdin (Ed.), *Encyclopedia of psychology* (Vol. 7, pp. 353–355). Washington, DC, New York: American Psychological Association.
- Cacioppo, J. T., Crites, S., Berntson, G., & Coles, M. (1993). If attitudes affect how stimuli are processed, should they not affect the event-related brain potential? *Psychological Science*, 4, 108–112.
- Cacioppo, J. T., Crites, S., Gardner, W., & Berntson, G. (1994). Bioelectrical echoes from evaluative categorizations: I. A late positive brain potential that varies as a function of trait negativity and extremity. *Journal of Personality and Social Psychology*, 67, 115–125.
- Cacioppo, J. T., & Petty, R. E. (1981). Electromyograms as measures of extent and affectivity of information processing. *American Psychologist*, 36, 441–456.
- Cacioppo, J. T., & Petty, R. E. (Eds.) (1983). *Social psychophysiology: A sourcebook*. New York: Guilford.
- Cacioppo, J. T., Petty, R. E., Losch, M., & Kim, H. (1986). Electromyographic activity over facial muscle regions can differentiate the valence and intensity of affective reactions. *Journal of Personality and Social Psychology*, 50, 260–268.
- Cantor, N., & Mischel, W. (1977). Traits as prototypes: Effects on recognition memory. *Journal of Personality and Social Psychology*, 35, 38–48.
- Cattaneo, Z., Mattavelli, G., Platania, E., & Papagno, C. (2011). The role of the prefrontal cortex in controlling gender-stereotypical associations: A TMS investigation. *Neuroimage*, 56, 1839–1846.
- Chartrand, T. L., & Bargh, J. A. (1996). Automatic activation of impression formation and memorization goals: Nonconscious goal priming reproduces effects of explicit task instructions. *Journal of Personality and Social Psychology*, 71, 464–478.
- Conrey, F., Sherman, J., Gawronski, B., Hugenberg, K., & Groom, C. (2005). Separating multiple processes in implicit social cognition: The Quad-Model of implicit task performance. *Journal of Personality and Social Psychology*, 89, 469–487.
- Craik, F., Moroz, T., Moscovitch, M., Stuss, D., Winocur, G., Tulving, E., & Kapur, S. (1999). In search of the self: A positron emission tomography study. *Psychological Science*, 10, 26–34.
- Cunningham, S. J., & Macrae, C. N. (2011). The colour of gender stereotyping. *British Journal of Psychology*, 102, 598–614.
- Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, 56, 680–690.
- Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. L. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology*, 48, 1267–1278.
- Dienstbier, R. A. (1989). Arousal and physiological toughness: Implications for mental and physical health. *Psychological Review*, 96, 84–100.
- Donaldson, W., & Glathe, H. (1970). Signal-detection analysis of recall and recognition memory. *Canadian Journal of Psychology*, 24, 42–56.
- Dovidio, J. F., & Fazio, R. H. (1992). New technologies for the direct and indirect assessment of attitudes. In J. M. Tanur (Ed.), *Questions about questions: Inquiries into the cognitive bases of surveys* (pp. 204–237). New York: Russell Sage Foundation.

- Ekman, P. (1993). Facial expression of emotion. *American Psychologist*, 48, 384–392.
- Ekman, P. (2007). *Emotions revealed* (2nd ed.). New York: Holt.
- Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology*, 69, 1013–1027.
- Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, 50, 229–238.
- Fiske, S. T. (1980). Attention and weight in person perception: The impact of extreme and negative behavior. *Journal of Personality and Social Psychology*, 38, 889–906.
- Ekman, P., & Friesen, W. V. (1978). *Facial action coding system: A technique for the measurement of facial movement*. Palo Alto, CA: Consulting Psychologists Press.
- Gilbert, D. T., & Hixon, J. G. (1991). The trouble of thinking: Activation and application of stereotypic beliefs. *Journal of Personality and Social Psychology*, 60, 509–517.
- Gilbert, D. T., & Osborne, R. E. (1989). Thinking backward: Some curable and incurable consequences of cognitive busyness. *Journal of Personality and Social Psychology*, 57, 940–949.
- Gilbert, D. T., Pelham, B. W., & Krull, D. S. (1988). On cognitive busyness: When person perceivers meet persons perceived. *Journal of Personality and Social Psychology*, 54, 733–740.
- Greenwald, A. G., McGhee, & Schwartz, J. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74, 1464–1480.
- Guglielmi, R. S. (1999). Psychophysiological assessment of prejudice: Past research, current status, and future directions. *Personality and Social Psychology Review*, 3, 123–157.
- Hamilton, D. L., Katz, L. B., & Leirer, V. O. (1980). Organizational processes in impression formation. In R. Hastie, T. Ostrom, E. Ebbesen, R. Wyer, D. Hamilton, & D. Carlston (Eds.), *Person memory: The cognitive basis of social perception* (pp. 121–153). Hillsdale, NJ: Lawrence Erlbaum.
- Hamilton, D. L., & Stroessner, S. J. (2021). *Social cognition: Understanding people and events*. Thousand Oaks, CA: Sage.
- Higgins, E. T., Rholes, W. S., & Jones, C. R. (1977). Category accessibility and impression formation. *Journal of Experimental Social Psychology*, 13, 141–154.
- Hoffmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between implicit association test and explicit self-report measures. *Personality and Social Psychology Bulletin*, 31, 1369–1385.
- Kidder, L. H., & Campbell, D. T. (1970). The indirect testing of social attitudes. In G. Summers (Ed.), *Attitude measurement* (pp. 333–385). Chicago, IL: Rand-McNally.
- Lane, R. D., Reiman, E., Ahern, G., Schwartz, G., & Davidson, R. (1997). Neuroanatomical correlates of happiness, sadness, and disgust. *American Journal of Psychiatry*, 154, 926–933.
- Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1990). Emotion, attention, and the startle reflex. *Psychological Review*, 97, 377–395.
- Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1992). A motivational analysis of emotion: Reflex-cortex connections. *Psychological Science*, 3, 44–49.
- Logan, G. D. (1980). Attention and automaticity in Stroop and priming tasks: Theory and data. *Cognitive Psychology*, 12, 523–553.
- MacLeod, C. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, 109, 163–203.
- Macrae, C. N., Bodenhausen, G. V., & Milne, A. B. (1995). The dissection of selection in person perception: Inhibitory processes in social stereotyping. *Journal of Personality and Social Psychology*, 69, 397–407.
- McArthur, L. Z., & Ginsberg, E. (1981). Causal attribution to salient stimuli: An investigation of visual fixation mediators. *Personality and Social Psychology Bulletin*, 7, 547–553.
- Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, 106, 226–254.
- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 231–259.
- Norman, G. J., Hawkey, L. C., Luhmann, M., Cacioppo, J. T., & Berntson, G. G. (2013). Social neuroscience and the modern synthesis of social and biological levels of analysis. In D. D. Franks & J. H. Turner (Eds.), *Handbook of neurosociology* (pp. 67–81). New York: Springer Science + Business Media.

- Olson, J., & Zanna, M. (1979). A new look at selective exposure. *Journal of Experimental Social Psychology*, 15, 1–15.
- Ostrom, T. M., Pryor, J. B., & Simpson, D. D. (1981). The organization of social information. In E. T. Higgins, C. P. Herman, & M. Zanna (Eds.), *Social cognition: The Ontario symposium* (Vol. 1, pp. 3–38). Hillsdale, NJ: Lawrence Erlbaum.
- Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology*, 105, 171–192.
- Parker, S., Clarke, C., Moniz-Cook, E., & Gardiner, E. (2012). The influence of 'cognitive busyness' on causal attributions of challenging behaviour in dementia: A preliminary experimental study. *Aging & Mental Health*, 16, 836–844.
- Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology*, 81, 181–192.
- Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change*. New York: Springer-Verlag.
- Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience*, 12, 729–738.
- Pratto, F., & John, O. P. (1991). Automatic vigilance: The attention-grabbing power of negative social information. *Journal of Personality and Social Psychology*, 61, 380–391.
- Pryor, J. B., & Ostrom, T. M. (1981). The cognitive organization of social information: A converging-operations approach. *Journal of Personality and Social Psychology*, 41, 628–641.
- Quadflieg, S., & Mcrae, C. N. (2011). Stereotypes and stereotyping: What's the brain got to do with it? *European Review of Social Psychology*, 22, 215–273.
- Robinson, R. G. (1995). Mapping brain activity associated with emotion. *American Journal of Psychiatry*, 152, 327–329.
- Roener, D. L., Thompson, C. P., & Brown, S. C. (1971). Comparison of measures for the estimation of clustering in free recall. *Psychological Bulletin*, 76, 45–48.
- Schimmack, U. (2021). Invalid claims about the validity of implicit association tests by prisoners of the implicit social-cognition paradigm. *Perspectives on Psychological Science*, 16, 435–442.
- Srull, T. K., & Wyer, R. S., Jr. (1979). The role of category accessibility in the interpretation of information about persons: Some determinants and implications. *Journal of Personality and Social Psychology*, 37, 1660–1672.
- Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18, 643–662.
- Taylor, S. E., & Fiske, S. T. (1981). Getting inside the head: Methodologies for process analysis in attribution and social cognition. In J. Harvey, W. Ickes, & R. Kidd (Eds.), *New directions in attribution research* (Vol. 3, pp. 459–524). Hillsdale, NJ: Erlbaum.
- Taylor, S. E., Fiske, S. T., Etcoff, N., & Ruderman, A. (1978). The categorical and contextual bases of person memory and stereotyping. *Journal of Personality and Social Psychology*, 36, 778–793.
- Tomaka, J., & Blascovich, J. (1994). Effects of justice beliefs on cognitive appraisal of and subjective, physiological, and behavioral responses to potential stress. *Journal of Personality and Social Psychology*, 67, 732–740.
- Tomaka, J., Blascovich, J., Kelsey, R. M., & Leitten, C. L. (1993). Subjective, physiological, and behavioral effects of threat and challenge appraisal. *Journal of Personality and Social Psychology*, 65, 248–260.
- Van Bavel, J. J., & Cunningham, W. A. (2010). A social neuroscience approach to self and social categorisation: A new look at an old issue. *European Review of Social Psychology*, 21, 237–284.
- Vanman, E. J., Paul, B. Y., Ito, T. A., & Miller, N. (1997). The modern face of prejudice and structural features that moderate the effect of cooperation on affect. *Journal of Personality and Social Psychology*, 73, 941–959.
- Willingham, D. T., & Dunn, E. W. (2003). What neuroimaging and brain localization can do, cannot do, and should not do for social psychology. *Journal of Personality and Social Psychology*, 85, 662–671.

19

METHODS FOR ASSESSING DYADS AND GROUPS

The preceding chapters on measurement and social cognition dealt with methods for assessing characteristics and behaviors of individual persons. **Monadic variables** are characteristics of single persons (e.g., age). A person's attitude toward abortion, for instance, is considered monadic because it refers only to that specific individual's own attitude. In many areas of social science, however, we are concerned with persons who interact in dyads (pairs) or groups. In this case, we are not assessing properties of the individuals separately, but the nature, relationship, structure, process, or outcomes of the dyad or group as a unit. **Dyadic variables** are characteristics of the relationship between two persons or their combined outcomes (e.g., level of warmth of their interaction). **Group variables** refer to characteristics of a set of three or more persons (e.g., total number of widgets created by a workteam). In this chapter, we consider the assessment of group or dyadic variables. Measures of behaviors or attitudes of interacting persons are a special case because the assessments taken from each of the group members are *interdependent*, insofar as each group member's score is expected to be related to or influenced by that of the other member(s).

Dyadic Designs

A common vocabulary is used to describe members in a dyadic design (Cook & Kenny, 2005; Kenny & Lederman, 2010). The **actor** is the person who perceives, records, and provides information or a score regarding an interpersonal interaction with a partner in a dyad. The **partner** is the person with whom the actor is interacting during a particular dyadic exchange. Consider the following examples of three different studies of interacting dyads:

★ Dyad 1: Tom & Peter

- a How much Tom likes Peter using a 1- to 10-point scale.
- b How much Peter likes Tom using a 1- to 10-point scale.

★ Dyad 2: Dick and Paul

- a How many times Dick smiles when interacting with Paul.
- b How many times Paul smiles when interacting with Dick.

★ Dyad 3: Harry and Mary

- a How intelligent Harry thinks Mary is.
- b How intelligent Mary thinks Harry is.

In each of these three dyadic interactions, for part (a), the actors are Tom, Dick, and Harry, with Peter, Paul, and Mary being their partners, respectively.¹ For part (b), the roles of actor and partner are reversed. The person who provides the observation or score for the measure is defined as the actor. Thus, if both members of a dyad provide scores, then both members serve as actors as well as partners to each other. The actor's score on a measure is influenced not only by the actor's own characteristics, but also by those of the partner. For instance, Tom's liking of Peter is in part a consequence of something about Tom (e.g., his attitudes and behaviors), but it also is influenced by characteristics of Peter (e.g., his attitudes and behavior), and by the nature of the relationship between the two of them. All three of these examples, then, involve dyadic designs, because they are concerned with *two persons and their interaction*.

Dyadic variables may be classified into three types (Kenny et al., 2006): within-dyad variables, between-dyad variables, and mixed variables. Examples of the three types are shown in Table 19.1, in a scenario involving brother-sister pairs. A *within-dyad variable* is one in which variations in scores are found within members of each dyad, but the same total score is shown from dyad to dyad. An example is a variable representing a game in which each brother and sister pair competed to see who could win the most of ten marshmallows (Table 19.1). In this case, the total number of marshmallows is the same (10) for every dyad, but the distribution within dyads varies.

A *between-dyad variable* is represented by a measure in which the same score is assigned to both members of the dyad, with variations in scores exhibited from dyad to dyad. Continuing with the brother-sister scenario, such a measure might involve how many toys each pair shares together at home (Table 19.1). Sharing is a relational concept, so the distinction between actor and partner is unnecessary, and members of each pair receive the same (dyadic) score. Other examples of between-dyad variables include the physical distance between members of the dyad, the amount of mutual gaze, and the frequency of simultaneous speech of each pair – the same value serves for both members of the dyad as it describes the pair as a unit.

In the final type of dyadic measure, a *mixed variable* exhibits variations in scores within dyads and between dyads. For example, both members of the brother-sister pair report on their personal attitudes toward their mother, using a rating scale from 1 to 10. With such a measure, obtained scores could differ within each brother-sister pair and across sibling pairs as well.

TABLE 19.1 Examples of the three types of dyadic variables in a sample of brother-sister pairs

Pair	<i>Within-dyad variable</i> (# of marshmallows won out of ten)		<i>Between-dyad variable</i> (# of toys shared at home)		<i>Mixed variable</i> (attitudes toward mother)	
	Brother	Sister	Brother	Sister	Brother	Sister
Dyad 1	6	4	3	3	7	3
Dyad 2	3	7	4	4	10	8
Dyad 3	4	6	6	6	8	9
Dyad 4	9	1	1	1	5	8
Dyad 5	5	5	8	8	2	2

Dyadic research may vary in the methods used to assess characteristics of the dyads. If conducted as observational research, the investigator or coder might carefully watch the dyadic interaction and code for relevant variables. Behavioral observations could take place in a systematically controlled laboratory setting or in a naturalistic setting (see Chapter 13). Using self-report measures to assess attitudes or feelings after the interaction, measures might concern the actor (“Rate how positively you feel about yourself”) or the respective partner (“Rate how positively you feel about the other person”), or each person’s assessment of the positivity of the interaction itself.

Some studies might appear to be dyadic, but actually are not. An investigation is not considered dyadic if data are collected exclusively from only one member of the pair. This happens when one of the two persons interacting is a confederate or an accomplice of the experimenter. In studies involving confederates, there are multiple actors (participants), but usually only a single partner (the confederate). When each of the actors separately interacts with the same partner, the data from each actor can be considered monadic, particularly if the confederate’s behavior has been scripted by the researcher.² For example, consider research that measures the degree of self-esteem reported by participants after a confederate expresses a degrading insult. In such a study, the confederate is constant across all dyads. As only one member of each dyad provides a score, it is not truly a dyadic design. In this chapter, we will focus on the study of truly dyadic, not monadic, measures.

Deriving Dyadic and Group-Level Variables

Some variables of dyadic or group properties involve assessments of the interacting members in a dyad or group. For instance, the level of interpersonal attraction between two persons may be scored by an observer who watches an interaction between the pair and then rates that interaction in terms of degree of warmth, engagement, and mutuality expressed by the pair. Similarly, group cohesion (the group-level counterpart to interpersonal attraction) can be rated by observers of group process considering the group as a whole, or the performance of a group on a collective task may be assessed by evaluating the overall group product independent of the contributions of individual members.

Sometimes, overall scores for each dyad or group are derived from measures computed from each of the component members. The overall score is the sum or mean of the scores of the constituting members. For instance, the average attraction between Peter and Tom may be the mean (which is simply the sum divided by two) of the degree of liking that Peter expresses for Tom and the liking that Tom expresses for Peter. At the group level, cohesion may be the mean of all the group members’ expressed liking for the group as a whole.

In other cases, the dyadic analysis is based on the degree of difference found between the ratings or behaviors of the paired members. For instance, attitude dissimilarity is usually measured as the degree of difference between the attitude expressed by person A of the pair and the attitude expressed by person B. The physical distance between two persons also is a dyadic variable, because it is based on the distance between the physical location of the individuals. For example, we might record how far apart two participants stand from each other when they are having a conversation. At the group level, belief similarity or dissimilarity is assessed by some measure of the variability of the attitude scores of individual members. A smaller variability reflects greater similarity (consensus) among members.

Although dyadic variables may involve simple mean scores of the component members, group variables sometimes can be considerably more complex. For instance, designs containing multiple groups (to be discussed later in this chapter) require statistical procedures to assess the contribution of variables at the participant level and at the group level.

Dyadic and group-level variables may be derived from individuals, but sometimes they are meaningful only at the level of the dyad or group. Group size provides a concrete example of what we mean here. Size is measured by counting individuals who compose the group – each individual contributes to this measure at the group level. However, size as a conceptual variable exists only as a property of each group, not in the individual bodies that constitute each group. Dyadic similarity is another example. Although individual A has an attitude that can be measured monadically, as does individual B, the distance between the attitudes of the members of the pair is a characteristic at the dyadic level, and not meaningful at the participant level.

Measuring Reciprocity or Mutuality

When working with dyadic variables, several issues arise that do not occur when working with monadic ones. When one has scores from both members of the dyad – so that each person serves as both actor and partner – it is possible that the two scores are correlated. The *degree of reciprocity* (or correspondence) is the extent that scores provided by both members to the same measure are correlated and is usually of theoretical importance in dyadic research. Let us consider some examples:

- In a study of randomly paired couples (Walster et al., 1966), the investigators were interested in discovering the degree of reciprocity of attraction between the individuals who constituted the dating couples. That is, they were studying to what extent the members of the pairs mutually liked or mutually disliked each other.
- Numerous theorists (e.g., Desjarlais et al., 2020; Sprecher et al., 2013) have discussed the conditions that affect the degree of self-disclosure that occurs in two-person relationships. Of particular interest is whether disclosure from one person stimulates comparable disclosure from the other member of the pair. It is then theoretically useful to measure the degree of reciprocity in self-disclosure.
- Communication scientists have speculated that in communicative social interactions, one person will tend to be dominant and the other submissive. If we take the number of interruptions as a measure of domination, we may want to see if in dyads there is one person who interrupts the other repeatedly while the other rarely interrupts. In this case, the behavior is not reciprocated, but rather compensated, and the predicted correlation between measures is negative rather than positive.

These illustrations demonstrate that it is essential for many research questions to be studied in such a manner to allow for the investigation of reciprocity. The measure of reciprocity depends upon whether the dyad is symmetric or asymmetric. In a **symmetric dyad**, the members of each pair are indistinguishable by a relevant feature, such as their respective role or status in the relationship. Examples of symmetric dyads are friends (friend X and friend X'), identical twins (twin A-twin A') or roommates (roommate X-roommate X'). In an **asymmetric dyad**, the members of each pair are distinguishable by a relevant feature, such as their respective role or status in the relationship. For instance, members of married heterosexual couples are distinguishable (husband-wife), as are teacher-student, parent-child, brother-sister, and boss-employee. In the roommate example, if we studied roommates who are of different genders (female roommate X-male roommate Y), the dyads would be considered asymmetric. Conversely, for a gay couple (husband X and husband X') the dyad would be symmetric.

For asymmetric dyads, the degree of reciprocity can be measured by the Pearson correlation coefficient. So, for instance, to assess the degree to which members of a group of married couples

agree on how satisfied they are with their marriage, one simply correlates the satisfaction expressed by the husbands with that of their respective wives. A stronger positive correlation signifies greater correspondence within couples – if the husband is satisfied, so is the wife, and vice versa.

For symmetric dyads there is no rule on how to compute such a correlation because there is no obvious way to identify one individual's score as the "X" variable and the other as the "Y" variable. There are two common, yet faulty, solutions to this problem. The first is to find some way of distinguishing the two members of the dyad. For example, if the dyads are same-sex siblings, they can be distinguished by the relative age (younger-older) of each dyad's members. The second way is to distinguish the members within a dyad in some arbitrary fashion. For instance, one uses a coin flip to designate one individual as Person 1 and the other as Person 2 and repeating this process for each dyad pair in the sample. We then enter the score of Person 1 as the X variable and that of Person 2 as the Y variable and compute the correlation between the scores as the measure of reciprocity.

For symmetric dyads, neither of these two approaches is ideal. The first, which calls for the discovery of a relevant variable by which to distinguish people, is not always feasible because there may be no reasonable characteristic to distinguish members of the dyad. The second approach, making use of an arbitrary (or relatively meaningless) rule to distinguish members, also is misguided because minor variations in who occupies each role can influence the obtained results. The correct approach makes use of a specialized measure of association called the *intraclass correlation coefficient*, which is a statistic that compares the variance within dyads to the variance between dyads (Kenny et al., 2006). For dyads, the intraclass correlation, like an ordinary correlation coefficient, ranges from -1 to $+1$, with zero indicating no relation between the scores within dyads. The intraclass correlation is the most appropriate measure of degree of correspondence in scores for members within symmetric dyads.

In Table 19.2 we present scores from eight pairs of college roommates as an example. The scores indicate the extent to which an individual trusts his or her roommate. The intraclass correlation for these pairs of roommates is .43, which indicates that the roommates reciprocated their trust in one another – that is, if one member of a pair had high trust in the other, it is likely that his or her roommate would indicate relatively high trust also. A standard Pearson correlation coefficient on the data of Table 19.2 would reveal a slightly different result ($r = .51$). However, as noted, the Pearson r is susceptible to variations in the placement of the data into variable X or X'; Thus, if the scores of Pair A of the table were reversed, the Pearson r would be affected (now, $r = .48$), whereas the intraclass correlation would not. For this reason, the intraclass correlation is clearly preferable in analyzing symmetric dyads.

TABLE 19.2 Trust scores from eight pairs of college roommates

Pair	Scores	
	Roommate X	Roommate X'
A	7	10
B	8	8
C	7	7
D	8	7
E	7	8
F	7	8
G	4	6
H	6	6

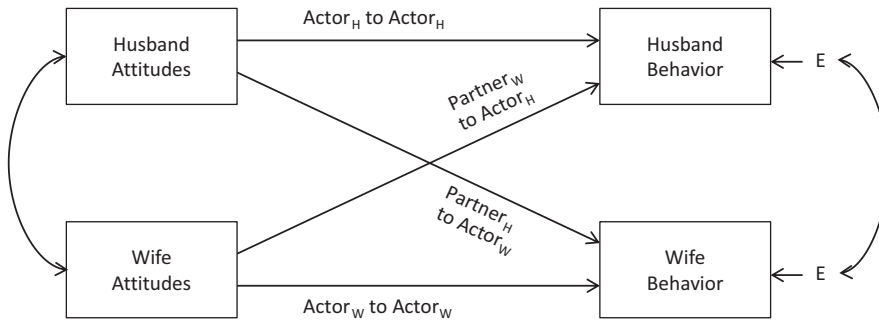


FIGURE 19.1 Actor-partner interdependence model.

Note: H = Husband, W = Wife.

The Pearson correlation and the intraclass correlation are used to assess the correspondence between scores on one variable collected from each member of the dyad. The **actor-partner interdependence model** is a multivariate framework for examining two or more variables from each member of the dyad to determine the correspondence and reciprocity of their interpersonal relationship (Cook & Kenny, 2005; Kenny & Ledermann, 2010). Typically, this framework is estimated using a multiple regression or a structural equation model (Chapter 11). These statistical techniques enable the researcher to statistically control, and therefore disentangle, the unique and simultaneous contribution of members of a dyad.

Although using the actor-partner interdependence model to examine symmetric dyads is possible, a correction is necessary to account for the fact that members are statistically indistinguishable (Kenny & Ledermann, 2010). Applying the model to asymmetric dyads is more straightforward. Depicted in Figure 19.1 is the actor-partner interdependence model for a sample of married heterosexual couples. Both members of each husband-wife pair provide scores on at least two variables (attitudes and behaviors) assessed at two points in time. (Although the model may be estimated with cross-sectional data, longitudinal collection of data enables the unraveling of temporal events.) Assessed at the initial round are measures of attitudes (e.g., level of trust), and at the later round are the measures of behaviors (e.g., number of chores performed). The scores of husband-wife dyads are expected to be interdependent: it is reasonable to assume that because of their intertwined lives, the scores of each member will be affected by many of the same household events. Thus, their scores are allowed to be correlated at each measurement round. The predictive path weights (one-directional arrows) are beta (β) weights, statistically controlling for the attitudinal variables of both members. Thus, the husband's behavior is a simultaneous function of his own and his wife's attitudes. Likewise, the wife's behavior is simultaneously predicted by her own as well as her husband's attitudes. Given that both husband and wife provide scores, they are actors as well as partners to each other.

Designs to Study Group Structures

Round Robin Designs

One problem with dyadic measurements is that it often is difficult to distinguish monadic and dyadic components contributing to the derived score. For instance, the overall attraction between Tom and Harry (their mean ratings of each other) may tell us something about the nature of the relationship between the two of them – something that is unique to them as a specific

dyad. However, the score at the dyadic level also reflects something about Tom and Harry as individuals. Perhaps Tom is the kind of person who expresses liking for everyone he meets. In that case, any dyad with Tom as a member would have a high attraction score, but this may be more a property of Tom than of the relationship with his partner. Even a physical distance score may be primarily determined by the behavior of one member. For example, Tom might have a dispositional tendency to stand in close proximity to all other people. When all we have is one dyad with Tom as a member, we cannot tease apart how much the overall attraction score is due to Tom alone and how much it reflects that particular dyad.

So far, we have assumed that each person under study is a member of one and only one dyad. If there are a total of 16 persons, as in the earlier example in Table 19.2, there would be a total of eight dyads. If the researcher wants to examine the interaction of each member paired with every other member in a group, it is necessary to employ a more complicated design. A **round robin design** is used to examine multiple combinations of dyads in a group, by creating all possible pairs of persons, with each member rating every other member. For instance, consider the now classic research by Newcomb (1961), who studied a group of 17 men living in the same university house. He asked each of these college men to rate how much he liked each of the others. With 17 respondents, the total number of dyads that can be formed (to compare the liking of each actor for each potential partner in the group) is calculated by the formula $N \times (N-1)/2$, in this case, $17 \times 16/2$, or 136 dyads.

Round robin designs are used infrequently in the social sciences, as they require all possible combinations of dyads. However, they are used on occasion. In research on interpersonal attraction, we have already mentioned that Newcomb used a round robin design. In their replication of Newcomb's (1961) study, Curry and Emerson (1970) used the same method. Round robin designs have also been used to study communication in monkeys (Miller et al., 1967), intertribal relations in East Africa (Brewer & Campbell, 1976), and defense mechanisms in person perception (Campbell et al., 1964).

There are two major advantages of collecting data using a round robin design. First, one obtains many more observations from each participant without increasing the number of participants needed. This gain in the number of comparisons lends added power to the statistical analysis of the collected data. Second, with a round robin design one can determine how a person rates or responds generally to others, and how that same person is rated and responded to by others. It then is possible to describe the score from a specific dyad (actor and partner) as being a function of (a) the set of actors being studied, (b) how the actor responds in general, (c) how others respond in general to the partner, and (d) how the actor uniquely responds to that specific partner. Such an analysis of two-person relationships has been termed the Social Relations Model by Kenny and LaVoie (1984) (see also Kashy & Kenny, 2000).

Table 19.3 presents an illustration of a round robin design. The numbers in the table represent the liking scores (on a scale from 1 to 15) of six persons toward each other. The rows in the table are the actors (they are reporting the score), and the columns are the partner. Thus, all actors also serve as partners in pairings. So, for instance, the table indicates that John likes Paul (with a score of 12 on the liking scale), whereas Paul's liking for John is only nine. We have left blank the entries in which the same person is both actor and partner to signify that those numbers are not gathered (the study is focused on interpersonal liking, and not how much participants like themselves).

By reading down each column of Table 19.3, we can see who is rated by the other members as being popular overall, and who is not. It appears that the members typically perceived Bill as the most popular person ($15 + 13 + 15 + 8 + 15 = 66$) and Phil the least popular ($10 + 4 + 9 + 7 + 4 = 34$). Reading across rows, we can see the person who generally likes the other members

TABLE 19.3 Round robin design

Actor (Person doing the rating)	Partner (Person being rated)					
	John	Paul	Mike	Bill	Dave	Phil
John		12	12	15	15	10
Paul	9		4	13	11	4
Mike	14	9		15	15	9
Bill	11	8	7		9	7
Dave	6	8	7	8		4
Phil	12	10	8	15	13	

the most and the one who likes the other members the least. It appears that John tends to like his group members the most ($12 + 12 + 15 + 15 + 10 = 48$), and that Dave likes them the least ($6 + 8 + 7 + 8 + 4 = 33$). We can view the general tendency to like (or dislike) others as a *response set*, a tendency to respond in a consistent fashion.

With a round robin design, it is possible to obtain a more refined measure of reciprocity than we can obtain from pair data. Consider what would happen if persons who were rated as popular in the group tended not to like the other persons in the group. If this were true, it would build a negative bias into the reciprocity correlation. For instance, for the data in Table 19.3, the intraclass correlation across the pairings of dyads is $-.38$. Once overall popularity and liking response sets are controlled, however, the correlation becomes $.65$. Thus, statistically controlling for, and thereby removing the effects of popularity and the liking response set gives a more accurate estimate of the intraclass correlation in a round robin design (Kenny & LaVoie, 1982).

To measure reciprocity in attraction more validly, we need to subtract from each score the extent to which the actor likes the other group members in general, and the extent to which the partner is liked by the others.³

Sociometric Analysis

Another approach related to the round robin design measures properties of groups such as cohesion and communication structure using a methodology developed in the 1930s by J. L. Moreno (1934) that became known as **sociometry**. Sociometry refers to the assessment of social acceptances and rejections of members in a group, by asking each member to indicate whether he or she likes or dislikes each of the other members. For example, Clore et al. (1978) asked each child in a summer camp to identify their best friend at the camp. They used these sociometric data to assess the degree to which children chose other children who were members of different ethnic groups.

In a sociometric design, there are two different kinds of choices that can be made by each member of a group. First, the participants can be asked to indicate whether they *accept* each of the group members. Examples of acceptance choices include liking, preferring to work with, preferring to be on the same team with, and the like. The second type of choice is *rejection*. Examples of rejection choices include disliking, rejection, and aversion of another member. The researcher can ask each participant to make choices regarding acceptances, rejections, or both, but in many studies, only acceptance choices are required. In that case, the investigator has the option of defining the number of choices a person can make. For instance, it is common to ask all respondents to choose their three best friends in a group (an “acceptance” choice). Although it seems more natural not to place any restrictions on the number of choices the participant can

make in this type of design, it is advisable to do so. If the number of choices is not set by the researcher, those who make many more choices than others in the group will have an inordinate influence on the resulting sociometric structures.

With sociometric data, we can represent participants' choices in what is called a *sociomatrix*, where each individual in the group is represented by a row and a corresponding column with all choices recorded in the appropriate cells. In Table 19.4 we provide an example of a sociomatrix for a group of six. In this example (because the number of group members is relatively small), each person in the group is asked to respond about each of the other group members, with acceptance choices represented by a "+," rejection choices by a "–," and non-choices by a "0." Each cell in the matrix represents the response that the individual in the corresponding row made for the individual in the column.

A sociomatrix employs a round robin design because data are generated for all possible pairs in the group. Although the design is round robin, the data are of a different type. The data in a sociomatrix involve categorical responses of acceptance or rejection, whereas in the round robin example presented in Table 19.3, quantitative ratings are used to assess degree of acceptance-rejection.

There are several interesting questions that can be addressed by analyzing the data of a sociomatrix (Terry, 2000). One type of question concerns whether the group is in *balance*. According to Heider's (1958) theory, if A likes B, and B likes C, then A should like C. Measures of the degree of balance in a sociomatrix of N participants were developed by Holland and Leinhardt (1978). From a sociomatrix it also is possible to measure the *popularity* of individuals within the group. One needs simply to count the number of acceptances received by each person, and subtract from this sum the number of rejections, if any. A person who receives a higher computed score is deemed more accepted by others. This simple computation strategy is valid when the number of choices to be made by each participant is set by the investigator. If respondents are free to choose (accept or reject) any number of other members, somewhat more complicated methods must be employed to assess popularity.

One can also determine from a sociomatrix the degree of *reciprocity* in the group. We begin by counting the number of pairs who reciprocate attraction (i.e., both members choose the other), and divide this quantity by the total number of pairs. The resulting number is the proportion of reciprocated pairs. This quantity should be compared to a baseline proportion of reciprocated pairs that would be expected if respondents' choices were completely random. One useful baseline index involves calculating the total number of choices made in the group, dividing this number by the total number of choices possible, and squaring the resulting proportion.

TABLE 19.4 Sociomatrix for a six-person group

Person Doing the Choosing	Person Being Chosen					
	A	B	C	D	E	F
A		+	+	0	–	–
B	+		0	–	0	–
C	0	0		+	–	–
D	–	0	+		0	–
E	–	–	0	0		+
F	–	0	0	0	+	

Note: + indicates acceptance, – indicates rejection, 0 indicates non-choice.

TABLE 19.5 A sociomatrix of fraternity member choices

Member Doing the Choosing	Member being chosen																	
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
A	–	+	+		+													
B			+		+	+												
C		+					+	+										
D		+	+		+													
E		+		+		+												
F		+					+				+							
G			+	+	+													
H									+	+			+					
I								+			+		+			+		
J								+	+				+					
K								+	+			+						
L								+		+			+					
M								+		+				+				
N									+				+					
O		+										+						
P															+		+	+
Q										+						+		+
R										+						+	+	

For example, if there were nine persons in a group, and each made three acceptances, the total number of acceptance choices made is 9×3 , or 27, and the total number of pairs is 72 (9×8). The baseline proportion of reciprocal choices thus is $(27/72)^2$, or .141. That is, if the choices were being made without any mutuality, 14% of the pairings would turn out to be matched acceptances by chance. This baseline proportion would be compared to the actual number of reciprocated choices to determine the extent of reciprocation (beyond chance levels) in the group under study.

From a sociomatrix we also can detect the presence of *cliques* or subgroupings that exist within a group. Table 19.5 presents an example of a sociomatrix of acceptance data from 18 members of a hypothetical fraternity. Each of the members was asked to state which three other members they liked the most. The matrix has been arranged to show the three clear subgroupings that were formed. The subgroupings are indicated in the table by boxes. These subgroupings were determined by a sociomatrix procedure called *block modeling* (White et al., 1976). (Note that Member “A” is a loner in that no others chose him among their three most liked, but he is included in the first subgrouping because all his choices fall in that group.)

A sociomatrix can be drawn as a figure to indicate who likes (acceptance choice) whom, and who does not. Such a drawing is called a *sociogram*. If person A likes person B, the two are connected by a solid arrow going from A to B, and vice versa if the choice is reciprocated. If rejection choices are made, they too are indicated by directional arrows, but acceptance choices are indicated by a plus, and rejection by a minus. We have taken the sociomatrix in Table 19.4 and redrawn it as a sociogram in Figure 19.2. Sociograms are more useful than sociomatrices if the number of respondents is less than ten. If we were to draw a sociogram for the sociomatrix in Table 19.5, the resulting diagram would be much too unwieldy to interpret.

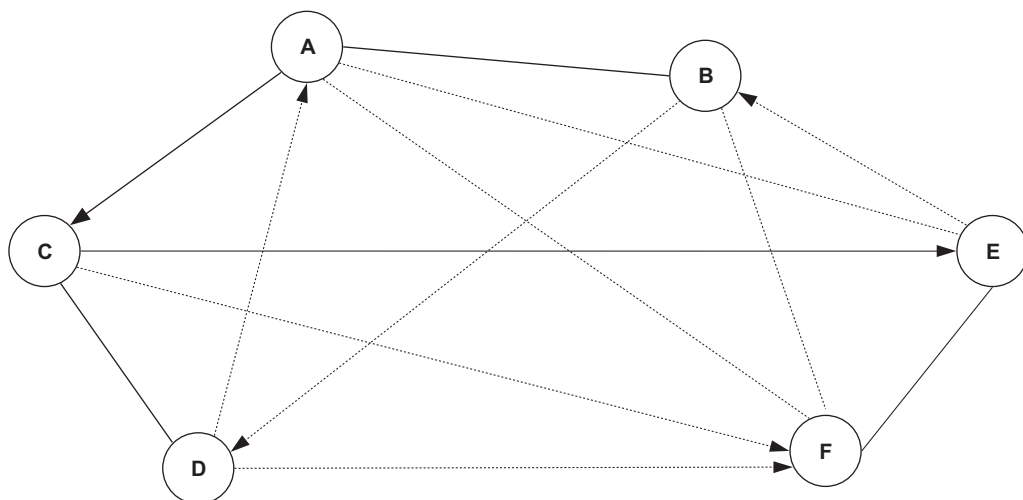


FIGURE 19.2 Sociogram of person relations presented in Table 19.4.

Social Network Analysis

Sociometric analysis is concerned with the affective ties among members of a social group using acceptance and rejection choices. Social network analysis is a related methodology concerning the communication links or connections among units (individuals or organizations) in a social system. Like sociometry, **social network analysis** uses dyadic relations as the basic unit of analysis, with the resulting data organized into matrix representations. However, the two methodologies differ in several ways. Whereas sociometry is usually applied to studies of relatively small, bounded groups, social network analysis often involves large social systems where the member units may be organizations or political units as well as individuals. The basic measure in a sociometric analysis is the acceptance or rejection relationships among group members, but the basic measure in a network analysis is some index of connectedness, communication, or exchange among component members. The specific kind of linkage being studied depends on the researchers' objectives and interests, but some common measures are exchanges of information or commodities (i.e., trade relations), presence or absence of telecommunication links, shared memberships (e.g., overlapping Boards of Directors), kinship relations, or collaborative efforts (e.g., co-authored publications). Data regarding the existence of linkages may be obtained from self-report questionnaires or interviews with the actors (or organizational representatives) or from archival sources such as publication citation indices, trade records, or IRS reports.

As with sociometry, the raw data on linkages for a social network analysis are compiled into an $N \times N$ matrix (where N is the number of members). When the linkage data are symmetrical (e.g., the presence of trade or kin relations), only the $N \times (N-1)/2$ elements below the diagonal in the matrix are needed for the database, but when measures are asymmetrical (e.g., A initiates communication with B, vs. B with A), the full matrix is used. Measures derived from this matrix data can relate to individual units or to the whole system (Knoke & Kuklinski, 1982; Scott, 1991; Wasserman & Faust, 1994) and statistical techniques often use advances in graph theory and other complex mathematical models (Robins, 2013). Measures reported by members include number of direct linkages to other units in the system, closeness (ease of interaction with others), and centrality. Other measures describe the nature of the linkages themselves,

such as stability, symmetry, or strength. Finally, other analyses focus on characteristics of the network overall, such as measures of degree of connectedness (saturation), centralization, and other structural features. Because of the versatility of this methodology, it is used widely across the social science disciplines interested in the study of social systems, including sociology, communication science, anthropology, science and technology studies, and organizational science (e.g., see Abbott et al., 2012; Cillissen, 2009; Eveland & Kleinman, 2013; Meisel et al., 2013; Monge, 1987; Trotter, 2000).

Designs to Study Multiple Groups

The study of dyads or a single group is an essential feature of almost all the social sciences. Also essential is the study of the behavior of many groups, each containing individual members. In round robin designs and sociometry, we have considered memberships in a single group or network. Now we shift attention to designs involving multiple groups. Studying variations across individuals who are nested within different groups calls for a methodological perspective different from that adopted when investigating individuals within a single group.

Nested Designs

The major design used by social scientists to study people interacting in groups and differences across the groups is known as multilevel modeling (MLM), which was discussed in Chapter 11. Usually in these designs, each person is a member of one and only one group. Consider a study of elementary school children in which the researcher has 300 students available to serve as participants. These children are found in ten different classrooms in the school. Each student, of course, is a member of one classroom, and this raises a potential problem for the researcher, because the scores of participants who are the members of the same group usually are more similar than the scores of those who are members of different groups. This similarity of scores happens because of conformity, social norms, modeling, and other social influence factors that operate when members within the same group interact. Even when the members of a group do not interact with each other directly, they may be exposed to common experiences or events that have similar effects on all of them. For instance, students in the same classroom not only interact with each other, they also share a common teacher and a common environmental setting (e.g., same grade level, size of classroom) that may influence them all. Because of these shared social factors, the scores of people in the same group are likely to be interdependent, making the units statistically *nonindependent*. As will be seen, this nonindependence must be considered in the data analysis.

With data from the hierarchically nested design of this type (i.e., participants' scores are nested within classrooms), one can use either the person or the group as the unit of analysis. In this classroom example, if person is the unit of analysis, correlations between variables and other statistics are computed for the sample of 300 persons. When the group is the unit of analysis, variables at the group level serve as the data in statistical analysis. In this case, the sample size would be ten, because ten classrooms (groups) were used, and thus, the group is the unit of analysis (vs. 300 if person was the unit).

When scores of members of the same group are statistically independent (e.g., students in the same classroom are not any more similar than they are to students of another classroom), then the individual can be used as the unit of analysis. This is desirable because it permits a more powerful statistical analysis of data, and statistically significant relations are more likely to be revealed if more, rather than fewer, units are employed in the analysis. However, if the scores

of persons within each group are not statistically independent, then group analytic techniques such as multilevel modeling should be used to simultaneously estimate effects at the grouping level (e.g., classrooms) and the participant level (students). A general approach to the assessment of dependence in dyadic and group data is also given by Kenny et al. (2006) and Sadler and Judd (2001).

Other Group Designs

Used much less frequently than the hierarchical nested models for studying participants who are nested in different groups are two types of designs for groups with small numbers of participants. In the *generations design*, one begins with a single group of a given size – say, three units, or actors. One makes the appropriate measurements on persons A, B, and C. Then person A is removed from the group and a new individual, Person D, is added. After measurements are taken, Person B is removed, and Person E is added. The process of removing a person and adding another is continued, and eventually the group consists of no original members. This design mimics in a laboratory the life and death replacements of cultures. The design was first used by Jacobs and Campbell (1961) to study the persistence of norms in groups. The researchers inserted a confederate into an original group of five persons, and the confederate influenced the group members to give a certain response. The confederate was then removed, a new (naïve) participant took his place, and the behavior of the group was studied. Jacobs and Campbell were able to show that adherence to the norm set by the confederate was still evident many “generations” after the confederate, and all the original group members had been removed from the group. Although used infrequently, the generations design can be used to answer some questions much more effectively than many of the more popular alternatives.

Another design that can be used in small group research is the *rotation design*. In the rotation design we begin with a set of, say, nine persons. Then, each person is placed in a group with each other person once and only once. With nine persons and a group size of three, each person would serve in four different groups. If we denote the persons as one through nine, we would have a total of 12 group combinations as depicted in Table 19.6.

At each time point, three different groups, labeled A, B, and C, are formed. No two persons are in the same group more than once. For instance, Person 3 is in a group with Persons 1 and 2 at time 1, with Persons 4 and 8 at time 2, with Persons 5 and 7 at time 3, and with Persons 6 and 9 at time 4. The person is in a group with each of the eight others once and only once, thus satisfying the requirements of the rotation design. Essentially, this is an extension of the round robin design from combinations of dyad pairings to combinations of group placement.

Barnlund (1962) used a rotation design to study whether the same person emerged as a leader, regardless of the group. He created groups of five people, and each person interacted in six

TABLE 19.6 The rotation design

Group	Time			
	I	II	III	IV
A	P1, P2, P3	P1, P5, P9	P1, P6, P8	P1, P4, P7
B	P4, P5, P6	P2, P6, P7	P2, P4, P9	P2, P5, P8
C	P7 P8, P9	P3, P4, P8	P3, P5, P7	P3, P6, P9

Note: P = person.

sessions (corresponding to six time points). No two persons were in the same group more than once. At each of the six sessions, the groups worked on a different type of task: artistic, construction, intellectual, coordination, literary, or social. Barnlund found that there was a fair degree of stability in leadership: Participants who sought a leadership role tended to do so across sessions, regardless of the task. The rotation design, though used infrequently, can be used to examine the stability of a person's behavior across different groups and situations. Kenny and Hallmark (1992) have provided software for analysis of rotation designs.

Measuring Group Process and Outcomes

Most of the methods for assessing group-level variables that we have discussed thus far involve measuring structural features of the grouping unit. Many of these measures can be obtained without observing the group members in interaction, through reports from group members or from archival records. Many of the more interesting and challenging issues in the study of groups arise from attempts to code and quantify the processes of group interaction itself, as well as the group performance, decisions, or products that arise from those processes. The study of group interaction process is a topic of interest in many social and behavior science disciplines, including social psychology, communication sciences, administrative sciences, sociology, education, and clinical psychology (McGrath & Altermatt, 2001). Consequently, the nature of such measures is as varied as the purposes and composition of groups themselves, so a complete cataloging of measures and data collection techniques is beyond the scope of this chapter (for overviews, see Kerr et al., 2000; McGrath, 1984; McGrath & Altermatt, 2001). However, we can provide some discussion of the most common methodological issues and paradigms that arise in these types of studies.

Interaction Processes

Process measures are sometimes derived from retrospective reports and ratings by members within each group after interaction has taken place. However, most studies of group process involve observation of the actual interactions among group members as they unfold in real time – either through direct observation or by videotaping the group in process and then analyzing the video record. In Chapter 13, we considered the most well-known observational system for classifying and recording interactive behaviors within groups – Bales' (1950, 1970) Interaction Process Analysis (IPA). The IPA and related observation systems (e.g., Borgatta & Crowther, 1965) provide a detailed record of what takes place during group interaction. Using the IPA generates categorical data representing the sequence of occurrence of interactive acts. Various methods of aggregating or summarizing the data from these records can be used to characterize the processes that have taken place in that group. For example, member contributions can be measured by aggregating each member's speaking turns during the group session and computing the proportion of total speaking turns for each.

Data from the IPA also permit quantitative summaries of the content of member contributions. For instance, Slater (1955) used IPA records to identify the roles that evolve among group members as they interact. He distinguished between two important types of functions that could be carried out by different group members – socioemotional functions (behaviors that help to maintain good relationships among group members) and task functions (behaviors that contribute to getting the group's job done). Subsequent research on group effectiveness has used these two variables, as well as other role measures derived from IPA data, to assess how well specific groups are functioning (Forsyth, 1990).

The frequency data generated from interaction process analyses can be analyzed using various statistical techniques for analyzing categorical data, such as Chi-square contingency analysis or loglinear and logit models (Argesti, 1996; Tian, 2013). However, the dynamic nature of group process often involves studying groups over time, which adds to the complexity of data analyses. In analyzing measures taken from the same interacting group at sequential points in time, we must consider *serial dependence* among behaviors over time. Just as measures taken from members within the same group are nonindependent (i.e., correlated with each other), so, too, the set of behaviors in a group are to some extent dependent on the behaviors that occurred just prior to it. For example, a group containing competitive members will likely remain competitive when assessed again at a later time point. Sequential analyses (e.g., Gottman & Roy, 1990) are used to assess the degree of serial dependence in a data set. (See Chapter 11 for further discussion of time-series analyses.)

Cognitive Processes in Groups

In addition to behavioral acts that can be observed during the course of interaction among members of a group, groups also have a cognitive life, i.e., the knowledge structures and information processing that are brought to bear to make group interaction and coordination possible. The study of group cognition has generated considerable interest in the social sciences (Theiner, 2013; Tindale et al., 2001; Wegner et al., 1991). For this purpose, novel methodologies have been developed to assess such things as how much and what information is shared among group members, how information is distributed, what group members know about each other's knowledge and skills, and what kind of "mental models" group members have about their task and structure.

The distribution of knowledge in a group may be analyzed by coding the contents of each member's knowledge prior to group interaction and assessing the degree of overlap between information held by one member and that of other members of the group (e.g., Kameda et al., 1997). The degree of overlap in knowledge prior to interaction can be compared with that obtained after the group interaction. The amount of shared information that members have at the outset of the group process can also be experimentally manipulated, as in the *hidden profile technique* developed by Stasser and Titus (1987, 2003). In this paradigm, members of four-person groups are instructed to engage in discussion for the purpose of coming to a consensus in selecting the best of three political candidates. Before discussion begins, each of the members is given a different set of information (some positive, some negative) about each of the three candidates. Some of the information is given to all four members (shared knowledge), and other information is uniquely given to just one of the members (unshared knowledge). During the discussion, if members opt to verbally contribute all the available information (shared + unshared), Candidate A is clearly the best choice. However, no one individual member of the group is aware that doing so would be beneficial. Thus, the group's decision (to choose Candidate A vs. B or C) will reflect how much the unshared information is brought up and used in the group discussion. Using this paradigm, researchers have found that shared knowledge is more likely to be discussed than is unshared knowledge, and as a result, groups often miss the "hidden profile" of information and fail to choose the best candidate.

In addition to the knowledge and information that individual group members bring to a task, they also develop *metacognitions* as the group interacts across time (Hinsz et al., 1997). At the group level, metacognition refers to the knowledge members possess about the knowledge and skills of each of the other group members, sometimes called *transactive memory* (Ren & Argote, 2011; Wegner, 1987), and about their understanding of the collective goals of the group

(sometimes called *shared mental models*, Klimoski & Mohammed, 1994). Tindale et al. (2001) provide a good illustration of group metacognition in their description of sports teams:

... For example, each of the nine members of a baseball team must have an understanding of the rules of the game and the roles for each player...for the team to work together. Thus, team players must have a mental model of the task (rules of the game) and the group (the roles of each player) ...to play effectively. However, this knowledge must be shared among the members in order for it to aid in team effectiveness. Two players who have different models of how to react in a given situation could each behave in ways that would interfere with the other's behavior.

(p. 18)

The importance of giving groups opportunity to develop shared mental models and transactive memory was demonstrated experimentally by Moreland et al. (1998). They found that when three-person groups were trained together (in a common session, without direct interaction among members) to perform a task of assembling a radio, they were subsequently able to perform the same task significantly better than groups who had been trained on the same task individually. Methods such as those used by Moreland and his colleagues are teaching group researchers much about the role of shared cognition in group process and coordination.

Computer-Mediated Group Interaction

Another innovation in the study of small groups is found in research on computer-mediated communication, which provides new ways of recording interactive behaviors and group process, as well as new questions to be addressed by group researchers (Bazarova & Yuan, 2013). Computers can be used as a tool for accomplishing group tasks, as when group members work together at a single computer or work at different computers with shared access – a process known as “groupwork” (e.g., McGrath & Hollingshead, 1994; Olson et al., 1994). But most often, computers are used by groups as a medium of communication among members who are not in face-to-face contact (e.g., Walther, 2012; Walther et al., 2010). Research comparing computer-mediated communication with face-to-face group interaction addresses interesting questions about the role of nonverbal cues in group process and interpersonal coordination, effects of status cues on rates of participation, and amount of information shared among group members (for reviews see Hollingshead & McGrath, 1995; Kiesler & Sproull, 1992). For example, the importance of paralinguistic and nonverbal communication to transactive memory in dyads was documented in experiments by Hollingshead (1998). In this research, dating couples worked together on a general knowledge test either through computer-mediated or face-to-face communication. Face-to-face couples performed significantly better than the computer-mediated couples, apparently because they were able to use nonverbal cues to assess which partner knew the correct answer. Couples who did not know each other prior to the experiment did not show this advantage, suggesting that the effective use of nonverbal information is part of a transactive memory that develops over time.

Computer-mediated groups also become more effective over time. In a longitudinal study comparing computer-mediated and face-to-face communication, Hollingshead et al. (1993) examined task performance of work teams, formed from an undergraduate psychology class, in which the task involved writing a joint paper on course materials. The work groups met weekly in two-hour lab sessions over the course of 13 weeks. Results showed that the computer-mediated groups had poorer task performance than face-to-face groups initially, but after three

weeks differences were no longer evidenced in the quality of papers produced by the two types of teams. Because of the role of learning in all forms of group process, the results of longitudinal studies of groups may show very different results from those obtained in cross-sectional studies at a single point in time. The dramatic increase in reliance on virtual interaction and computer-mediated team meetings that occurred during the COVID-19 pandemic provided social psychologists and organizational researchers a wealth of new data on the advantages and disadvantages of computer-mediated systems on group process and performance (e.g., Hacker et al., 2020; Karl et al., 2022).

Comparing Groups and Individual Performance

A long-standing issue in the study of groups is that of evaluating the “value-added” by having people work together in groups rather than as lone individuals. Some tasks clearly require the concerted effort of multiple individuals working together because the nature of the task itself is beyond the capabilities of any one person to accomplish (e.g., moving a large and heavy structure, fighting an invading force, getting to the moon). For many intellectual or physical tasks, however, it is not always clear whether groups of persons working cooperatively together produce more or better products than could be attained by having individuals work separately and then pooling the output of their individual efforts. To address this question, social researchers have developed the method of comparing performance (or products) of a real, interactive group, with the products of a **nominal group**, which is composed of the same number and types of people as the real groups, but the members work independently rather than in an interaction, and the group product is a combination of their individual performances or output.

One interesting example of the evaluation of group performance comes from the study of *brainstorming groups*. Brainstorming was initially developed by an advertising executive (Osborn, 1957) as a method for enhancing of creative idea generation through group interaction. The idea of a brainstorming group is that members throw out ideas in a freewheeling fashion, without evaluation or censure, building on ideas as they are generated. Based on the notion that interaction would both stimulate and inspire new ideas and combinations, Osborn made some very strong assumptions about the importance of group processes for creative output. For example, he claimed that “the average person can think up twice as many ideas when working with a group than when working alone” (Osborn, 1957, p. 229).

How has this claim held up to the results of systematic research? Alas for Osborn’s credibility, the results of several reviews of studies comparing interacting brainstorm groups with similarly instructed nominal groups consistently show that real groups generally produce fewer and poorer quality ideas than nominal comparison groups (Diehl & Stroebe, 1987; Mullen et al., 1991). Findings such as these have led group researchers to study the concept of *process loss* – those aspects of group interaction and coordination that inhibit or interfere with group production. For instance, in the case of brainstorming, the fact that group members must take turns talking may block production of ideas, and social psychological processes such as social comparison and evaluation apprehension may also inhibit output (Stroebe & Diehl, 1994). By reducing some of these influences, “process gains” may override “process losses,” producing conditions under which real groups outperform the nominal group (Dennis & Valacich, 1993).

Conclusion

The study of persons in dyads and groups allows for testing important hypotheses that could not be tested otherwise. Hypotheses concerning reciprocity, interpersonal attraction, and similarity

can best be tested with dyadic measures. To understand the processes operating in the study of groups and dyads, it is necessary to go beyond a naïve and intuitive analysis of data. Often, we must take a score obtained from a dyad or group and partition it into component parts – at the individual, dyadic, and group level – to interpret its meaning accurately. Such a partitioning was illustrated in our discussion of the round robin design, the measurement of attraction, and the analysis of scores of interacting groups. We need to divide a score into components that reflect the different processes that operate at different levels and that contribute to an observed score. It might seem that all the difficulties encountered when dealing with dyadic and group data diminish their attractiveness. We prefer to view these complications as interesting challenges that once confronted and conquered reveal fascinating and vital aspects of social interactions that could not be studied otherwise.

Notes

- 1 The terms actor and partner correspond to responder and stimulus, respectively, in a dyadic situation. Although we will generally use the terms actor and partner here, other terms can be used. For instance, Cronbach (1955) used the terms *judge* and *target*, and Swann (1984) uses the terms *perceiver* and *target*. In research on nonverbal communication, the terms *receiver* and *sender* as well as *decoder* and *encoder* correspond to actor and partner. We prefer the terms actor and partner because they are not specific to any one area of research.
- 2 Interactions that are computer-mediated (such as real or simulated “chat rooms”) lend themselves to this type of design. The partner in a computerized interaction is often a simulated creation of the experimenter, in which case the interest is in the behavior or responses of the single participant who is reacting to a preprogrammed stimulus person (e.g., Crano & Hannula-Bral, 1994). Sometimes, however, the researcher is studying real online interactions between two or more actual persons (e.g., McKenna & Bargh, 2000). In this case, the measurements would be dyadic or group-level variables.
- 3 We cannot take a simple average across rows to measure the response or a simple average across columns to measure popularity. More complicated methods are required, because the person does not interact with him or herself (Kenney et al., 1984; Warner et al., 1979).

References

- Abbott, K. M., Bettger, J., Hampton, K., & Kohler, H. (2012). Exploring the use of social network analysis to measure social integration among older adults in assisted living. *Family and Community Health: The Journal of Health Promotion & Maintenance*, 35, 322–333.
- Argesti, A. (1996). *An introduction to categorical data analysis*. New York: Wiley.
- Bales, R. F. (1950). *Interaction process analysis*. Cambridge: Addison-Wesley.
- Bales, R. F. (1970). *Personality and interpersonal behavior*. New York: Holt, Rinehart, and Winston.
- Barnlund, D. C. (1962). Consistency of emergent leadership with changing tasks and members. *Speech Monographs*, 29, 45–52.
- Bazarova, N. N., & Yuan, Y. (2013). Expertise recognition and influence in intercultural groups: Differences between face-to-face and computer-mediated communication. *Journal of Computer-Mediated Communication*, 18, 437–453.
- Borgatta, E. F., & Crowther, B. (1965). *A workbook for the study of social interaction processes*. Chicago, IL: Rand-McNally.
- Brewer, M. B., & Campbell, D. T. (1976). *Ethnocentrism and intergroup attitudes: East African evidence*. Beverly Hills, CA: Sage.
- Campbell, D. T., Miller, N., Lubetsky, J., & O’Connell, E. J. (1964). Varieties of projection in trait attribution. *Psychological Monographs*, 78 (issue 592).
- Cillissen, A. H. N. (2009). Sociometric methods. In K. H. Rubin, W. H. Bukowski, & B. Laursen (Eds.), *Handbook of peer interactions, relationships, and groups* (pp. 82–99). New York: Guilford.
- Clore, G. L., Bray, R. M., Itkin, S. M., & Murphey, P. (1978). Interracial attitudes and behavior at summer camp. *Journal of Personality and Social Psychology*, 36, 107–116.

- Cook, W. L., & Kenny, D. A. (2005). The actor-partner interdependence model – A model of bidirectional effects in developmental studies. *International Journal of Behavioral Development*, 20, 101–109.
- Crano, W. D., & Hannula-Bral, K. A. (1994). Context/categorization model of social influence: Minority and majority influence in the formation of a novel response norm. *Journal of Experimental Social Psychology*, 30, 247–276.
- Cronbach, L. J. (1955). Processes affecting scores on “understanding of others” and “assumed similarity.” *Psychological Bulletin*, 52, 177–193.
- Curry, T. J., & Emerson, R. M. (1970). Balance theory: A theory of interpersonal attraction? *Sociometry*, 33, 216–238.
- Dennis, A. R., & Valacich, J. S. (1993). Computer brainstorm: More heads are better than one. *Journal of Applied Psychology*, 78, 531–537.
- Desjarlais, M. (2020). Online self-disclosure: Opportunities for enriching existing friendships. In M. Desjarlais (Ed.), *The psychology and dynamics behind social media interactions* (pp. 1–27). Hersey, PA: Information Science Reference/IGI Global.
- Diehl, M., & Stroebe, W. (1987). Productivity loss in brainstorming groups: Toward the solution of a riddle. *Journal of Personality and Social Psychology*, 53, 497–509.
- Eveland, W. R., & Kleinman, S. B. (2013). Comparing general and political discussion networks within voluntary organizations using social network analysis. *Political Behavior*, 35, 65–87.
- Forsyth, D. R. (1990). *Group dynamics* (2nd ed.). Pacific Grove, CA: Brooks/Cole.
- Gottman, J. M., & Roy, A. K. (1990). *Sequential analysis: A guide for behavioral researchers*. New York: Cambridge University Press.
- Hacker, J., von Brocke, J., Handali, J., Otto, M., & Schneider, J. (2020). Virtually in this together: How web-conferencing systems enabled a new virtual togetherness during the COVID-19 crisis. *European Journal of Information Systems*, 29, 563–584.
- Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.
- Hinsz, V. B., Tindale, R. S., & Vollrath, D. A. (1997). The emerging conceptualization of groups as information processors. *Psychological Bulletin*, 121, 43–64.
- Holland, P. W., & Leinhardt, S. (1978). An omnibus test for social structure using triads. *Sociological Methods and Research*, 7, 227–256.
- Hollingshead, A. B. (1998). Retrieval processes in transactive memory systems. *Journal of Personality and Social Psychology*, 74, 659–671.
- Hollingshead, A. B., & McGrath, J. E. (1995). Computer-assisted groups: A critical review of the empirical research. In R. Guzzo & E. Salas (Eds.), *Team effectiveness and decision making in organizations* (pp. 46–78). San Francisco, CA: Jossey-Bass.
- Hollingshead, A. B., McGrath, J. E., & O'Connor, K. M. (1993). Group task performance and communication technology: A longitudinal study of computer-mediated versus face-to-face work groups. *Small Group Research*, 24, 307–333.
- Jacobs, R. C., & Campbell, D. T. (1961). The perpetuation of an arbitrary tradition through several generations of laboratory microculture. *Journal of Abnormal and Social Psychology*, 62, 649–658.
- Kameda, T., Ohtsubo, Y., & Takezawa, M. (1997). Centrality in sociocognitive network and social influence: An illustration in a group decision-making context. *Journal of Personality and Social Psychology*, 73, 296–309.
- Karl, K. A., Peluchetter, J. V., & Aghakhani, N. (2022). Virtual work meetings during the COVID-19: The good, bad, and ugly. *Small Group Research*, 53, 343–365.
- Kashy, D. A., & Kenny, D. A. (2000). The analysis of data from dyads and groups. In H. Reis & C. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 451–477). New York: Cambridge University Press.
- Kenny, D. A., & Hallmark, B. W. (1992). Rotation designs in leadership research. *The Leadership Quarterly*, 3, 25–41.
- Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*. New York: Guilford Press.
- Kenny, D. A., & La Voie, L. (1982). Reciprocity of attraction: A confirmed hypothesis. *Social Psychology Quarterly*, 45, 54–58.
- Kenny, D. A., & La Voie, L. (1984). The social relations model. *Advances in Experimental Social Psychology*, 18, 142–182.

- Kenny, D. A., & Ledermann, T. (2010). Detecting, measuring and testing dyadic patterns in the actor-partner interdependence model. *Journal of Family Psychology*, 24, 359–366.
- Kenny, D. A., Lord, R., & Garg, S. (1984). *A social relations model for peer ratings*. Unpublished paper, University of Connecticut.
- Kerr, N. L., Aronoff, J., & Messé, L. A. (2000). Methods of group research. In H. Reis & C. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 160–189). New York: Cambridge University Press.
- Kiesler, S., & Sproull, L. (1992). Group decision making and technology. *Organizational Behavior and Human Decision Processes*, 52, 96–123.
- Klimoski, R., & Mohammed, S. (1994). Team mental model: Construct or metaphor? *Journal of Management*, 20, 403–437.
- Knoke, D., & Kuklinski, J. H. (1982). *Network analysis*. Beverly Hills, CA: Sage.
- McGrath, J. E. (1984). *Groups: Interaction and performance*. Englewood Cliffs, NJ: Prentice-Hall.
- McGrath, J. E., & Altermatt, T. W. (2001). Observation and analysis of group interaction over time: Some methodological and strategic choices. In M. Hogg & R. S. Tindale (Eds.), *Blackwell handbook of social psychology: Group processes* (pp. 525–573). Oxford: Blackwell.
- McGrath, J. E., & Hollingshead, A. B. (1994). *Groups interacting with technology*. Thousand Oaks, CA: Sage.
- McKenna, K., & Bargh, J. A. (2000). Plan 9 from cyberspace: The implications of the internet for personality and social psychology. *Personality and Social Psychology Review*, 4, 57–75.
- Meisel, M. K., Clifton, A. D., MacKillop, J., Miller, J. D., Campbell, W., & Goodie, A. S. (2013). Ego-centric social network analysis of pathological gambling. *Addiction*, 108, 584–591.
- Miller, R. E., Caul, W. F., & Mirsky, I. A. (1967). Communication between feral and socially isolated monkeys. *Journal of Personality and Social Psychology*, 3, 231–239.
- Monge, P. R. (1987). The network level of analysis. In C. R. Berger & S. H. Chaffee (Eds.), *Handbook of communication science* (pp. 239–270). Newbury Park, CA: Sage.
- Moreland, R. L., Argote, L., & Krishnan, R. (1998). Training people to work in groups. In R. S. Tindale, L. Heath, J. Edwards, E. Posavac, F. Bryant, Y. Suarez-Balcazar, E. Henderson-King, & J. Myers (Eds.), *Social psychological applications to social issues: Applications of theory and research on groups* (Vol. 4, pp. 37–60). New York: Plenum Press.
- Moreno, J. L. (1934). *Who shall survive?* Washington, DC: Nervous and Mental Disease Monograph, No. 58.
- Mullen, B., Johnson, C., & Salas, E. (1991). Productivity loss in brainstorming groups: A meta-analytic integration. *Basic and Applied Social Psychology*, 12, 3–24.
- Newcomb, T. M. (1961). *The acquaintance process*. New York: Holt, Rinehart, & Winston.
- Olson, J. S., Olson, G., Storreston, M., & Carter, M. (1994). Groupwork close up: A comparison of the group design process with and without a simple group editor. *ACM Transactions on Information Systems*, 11, 321–348.
- Osborn, A. F. (1957). *Applied imagination*. New York: Scribners.
- Ren, Y., & Argote, L. (2011). Transactive memory systems 1985–2010: An integrative framework of key dimensions, antecedents, and consequences. *The Academy of Management Annals*, 5(1), 189–229.
- Robins, G. (2013). A tutorial on methods for the modeling and analysis of social network data. *Journal of Mathematical Psychology*, 57, 261–274.
- Sadler, M. S., & Judd, C. M. (2001). Overcoming dependent data: A guide to the analysis of group data. In M. Hogg & R. S. Tindale (Eds.), *Blackwell handbook of social psychology: Group processes* (pp. 497–524). Oxford: Blackwell.
- Scott, J. (1991). *Social network analysis: A handbook*. London: Sage.
- Slater, P. E. (1955). Role differentiation in small groups. *American Sociological Review*, 20, 300–310.
- Sprecher, S., Treger, S., & Wondra, J. D. (2013). Effects of self-disclosure role on liking, closeness, and other impressions in get acquainted interactions. *Journal of Social and Personal Relationships*, 30, 497–514.
- Stasser, G., & Titus, W. (1987). Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology*, 53, 81–93.
- Stasser, G., & Titus, W. (2003). Hidden Profiles: A brief history. *Psychological Inquiry*, 14, 304–313.
- Stroebe, W., & Diehl, M. (1994). Why groups are less effective than their members: On productivity losses in idea-generating groups. In W. Stroebe & M. Hewstone (Eds.), *European review of social psychology* (Vol. 5, pp. 271–303). London: Wiley.

- Swann, W. B. (1984). The quest for accuracy in person perception: A matter of pragmatics. *Psychological Review*, 91, 457–477.
- Terry, R. (2000). Recent advances in measurement theory and the use of sociometric techniques. In A. H. N. Cillessen & W. Bukowski (Eds.), *Recent advances in the measurement of acceptance and rejection in the peer system* (pp. 27–53). San Francisco, CA: Jossey-Bass.
- Theiner, G. (2013). Transactive memory systems: A mechanistic analysis of emergent group memory. *Review of Philosophy and Psychology*, 4, 65–89.
- Tian, G.-L. (2013). *Incomplete categorical data design: Non-randomized response techniques for sensitive questions in surveys*. Hoboken, NJ: Taylor & Francis.
- Tindale, R. S., Meisenhelder, H. M., Dykema-Engblade, A., & Hogg, M. A. (2001). Shared cognition in small groups. In M. Hogg & R. S. Tindale (Eds.), *Blackwell handbook of social psychology: Group processes* (pp. 1–30). Oxford: Blackwell.
- Trotter, R. T. II. (2000). Ethnography and network analysis: The study of social context in cultures and society. In G. L. Albrecht & R. Fitzpatrick (Eds.), *The handbook of social studies in health and medicine* (pp. 210–229). London: Sage.
- Walster, E., Aronson, V., Abrahams, D., & Rottman, L. (1966). Importance of physical attractiveness in dating behavior. *Journal of Personality and Social Psychology*, 4, 508–516.
- Walther, J. B. (2012). Interaction through technological lenses: Computer-mediated communication and language. *Journal of Language and Social Psychology*, 31, 397–414.516.
- Walther, J. B., Deandrea, D. C., & Tong, S. (2010). Computer-mediated communication versus vocal communication and the attenuation of pre-interaction impressions. *Media Psychology*, 13, 364–386.
- Warner, R. M., Kenny, D. A., & Stoto, M. (1979). A new round robin analysis of variance for social interaction data. *Journal of Personality and Social Psychology*, 37, 1742–1757.
- Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. New York: Cambridge University Press.
- Wegner, D. M. (1987). Transactive memory: A contemporary analysis of the group mind. In B. Mullen & G. Goethals (Eds.), *Theories of group behavior* (pp. 185–208). New York: Springer-Verlag.
- Wegner, D. M., Erber, R., & Raymond, P. (1991). Transactive memory in close relationships. *Journal of Personality and Social Psychology*, 61, 923–929.
- White, H., Boorman, S., & Breiger, R. (1976). Social structure from multiple networks, I: Block models of roles and positions. *American Journal of Sociology*, 81, 730–779.

PART IV

Concluding Perspectives



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

SYNTHESIZING RESEARCH RESULTS

Meta-Analysis

All sciences progress through the gradual accretion of knowledge. Although surely inspired, Einstein's classic formula relating energy to a function of mass and velocity was not developed out of thin air. His creative insight was based on a synthesis and extension of the available knowledge of the time (Clark, 1971; Kuhn, 1970). Over the years, researchers have depended on a similar process of integration and consolidation of the prior research literature to develop new insights, which could then inform the development of additional knowledge. A **primary study** is an original study reporting on the results of data analysis collected from a sample of participants (e.g., journal article, research report, thesis, dissertation). It serves as the principal unit of empirical knowledge. All the chapters covered to this point have described strategies for designing and conducting primary studies. However, we also have emphasized the importance of programmatic research and multiple studies directed at testing the same conceptual hypotheses. The current chapter is devoted to methods for integrating the findings from such cumulative research efforts.

Replicability of Findings

One hallmark of the scientific method is that research is conducted in such a way that results can be both observable and repeatable (see Chapter 1). Technically, replication means that if the exact methods used in an original primary study are applied in a new study, equivalent results would be obtained. In the past decade, some disciplines have gone through something of a crisis of confidence about the replicability of many of the findings in the published literature (Carey, 2015; Open Science Collaborative, 2015). The concern is due, in part, to the issue of a publication bias whereby prominent journals are more likely to publish reports of research in which the results are statistically significant, so studies with nonsignificant outcomes are less likely to appear in the published literature. This selection bias increases the probability of reporting "false positives" (i.e., Type I errors, as discussed in Chapter 2). Consequently, there has been a renewed focus in recent years on replication as the standard of credibility of reported research findings (Edlund et al., 2022).

Obviously, "exact" replication in using the same exact methods, participants, place, and time is never possible (Nosek et al., 2022; Stroebe & Strack, 2014). We think it is preferable to

think of a continuum of replication methods ranging from close duplication of the methods and operations of an original study with new participants in a new time and place at one end to replications that systematically change operations and measures to determine whether equivalent findings can be obtained with different methods at the other end. (See discussion of *conceptual replication* in Chapter 8.) Debates exist in the literature about the relative merits of “exact replication” versus “conceptual replication” (e.g., Crandall & Sherman, 2016), but we believe that there is a role for all types of replication efforts if properly conducted and interpreted.

The major problem with attempts at exact replication is how to interpret a “failure” to replicate. A duplication study is deemed to fail to replicate if it does not result in a statistically significant effect that is in the same direction as the original study. There may, however, be many reasons why the results of a replication do not match the original study even if the original effect was valid. For one thing, many empirical phenomena in social science research may be subject to “contextual sensitivity” (Van Bavel, 2016), meaning it can be influenced by minor changes in context, implementation of methods, or timing. Such sensitivity may indicate that a specific finding is not very robust, but it does not necessarily mean that the primary finding was a false positive.

Importantly, effects (or effect sizes) may not replicate exactly because of statistical variability (Stanley & Spence, 2014). Even a true effect that reaches the criterion for statistical significance in one study may be randomly smaller (or larger) in a duplicate study. One proposed solution to this problem is to conduct replications with very large sample sizes to have high power to consistently detect even small effects (Fiedler & Prager, 2018; LeBel et al., 2017). However, this makes replication very costly in terms of resources and participant pools at the expense of conducting new original research. A similar strategy is for different researchers to conduct multiple, small-scale replication studies across different laboratories to determine whether the finding replicates more frequently than expected by chance (Klein et al., 2014; 2018). This, too, is resource intensive, particularly when the focal study is just a single line of research. Overall, the value of replications must be weighed against their costs. Researchers must contend with trade-offs between conducting duplication studies and new original research (Miller & Ulrich, 2022). In our view, scientific progress is optimized when many different studies, including conceptual replications, are undertaken to test the same theoretical hypothesis, and then modern methods are employed to synthesize the results across these variations.

Meta-Analysis

The traditional constructive process of integrating research findings in a particular area of knowledge was based on the **narrative review**, a qualitative approach for summarizing and interpreting primary studies that have addressed the same research question. The time-honored narrative review, in which only words and sentences are used to describe and summarize a research area, has been supplemented in recent years by the development of methods of *quantitative reviews* and *syntheses*. The most widespread type is **meta-analysis**, a quantitative approach for summarizing and interpreting primary studies that have addressed the same research question. Although the traditional narrative review has served us well, its critics state that it is prone to shortcomings, including (a) the failure to review and include the existing knowledge base of primary studies comprehensively,¹ (b) the lack of clearly stated rules for the inclusion or exclusion of primary studies, and (c) the failure to apply statistical metrics that permit objectively combining findings across primary studies. A systematic and objective meta-analysis is intended to avoid these problems and thereby develop a comprehensive understanding of relations among variables. A major goal of meta-analysis is to use preexisting research to construct a secure foundation on which to build a knowledge base.

A meta-analysis offers a clearer path to understanding the true strength of relations between variables than does a single primary study. Meta-analysis had its beginnings many years ago, but the popularity of the technique was spurred by the refinement of more formalized statistics used to perform a synthesis (Cooper, 1979; Rosenthal, 1979, 1982).² Since that period, important and thought-provoking syntheses have been published to help popularize this approach. Meta-analysis represents a valuable tool, particularly when primary research on a topic has been investigated using different methods, an approach to research that we have championed throughout this book. A meta-analysis provides a way of assessing *construct validity* and *external validity* by pooling primary studies of the same hypothesis that may involve different methods and samples. When primary studies about the same topic or research question that are operationally or methodologically different produce similar results, the confidence in the construct validity of our findings is enhanced. Moreover, confidence in the external validity of a line of research is enhanced if different primary studies involving different participant samples – such as an undergraduate subject pool, participants recruited online, and people recruited at the mall – generate similar results.

An example will help highlight the role that meta-analysis can play in interpreting findings from methodological and sample variants of studies examining the same underlying hypothesis. Suppose a researcher was interested in reviewing the relation between physical attractiveness (independent variable) and liking (dependent variable) in the literature. Immediately, two relevant primary studies are located. The first study hypothesized that an attractive rather than an unattractive target would be rated more positively by participants. In this laboratory study, the researcher randomly assigned participants to a photograph of an attractive or unattractive person. Digitally altered photographs of the same original person served as the manipulation materials. Afterward, participants rated the person in the photograph using semantic differential items, such as pleasant to unpleasant, good to bad, and attractive to ugly. Participant ratings on these three items were internally consistent, so the researcher averaged them to compute an overall “liking” composite score. Statistical analysis disclosed that the physically attractive compared to unattractive target was rated significantly more positively (at $p < .05$).

Suppose the other primary study instead involved a field experiment in which the manipulation involved a confederate who was dressed to appear attractive or unattractive. In this study, a confederate individually approached strangers (participants) in a shopping mall and asked to borrow each stranger’s cell phone. The dependent measure of liking was operationally defined as participants’ speed of response (measured in seconds) in honoring the confederate’s request. The study used a small sample size and found that the attractive target compared to the unattractive target did not yield a statistically significant difference in this operationalization of liking ($p > .05$).

What conclusion may be drawn from these two primary studies with conflicting results? Does a relation exist between physical attractiveness and liking, as suggested in the statistically significant first study, or was this effect an artifact of the laboratory setting? Or, did the second study that produced a statistically nonsignificant result indicate that attractiveness does not increase liking, or is such a null finding due to lack of statistical power or operational definition of the liking variable (number of seconds for participants to lend the cellular phone to the confederate)? We find ourselves in a quandary because the two studies produced disparate conclusions about the connection between attractiveness and liking based on comparing the statistical significance of the results.

The problem with comparing primary studies on this basis is that statistical significance is determined by a variety of study characteristics (e.g., sample size and reliability of measures) that do not necessarily reflect upon the hypothesized association. All else being equal, a study

containing a larger sample size is more likely to attain statistical significance as it yields more stable estimates due to more units from the population being represented. Thus, probability values (p-values) obtained from significance testing are *not* directly comparable across primary studies, because they indicate *statistical significance*, not *practical significance*. To properly compare studies requires conversion to a common metric to denote the effect size obtained in each of the two studies. This is the basic insight and contribution of meta-analysis – to convert the statistical tests of primary studies that address the same hypothesis into the same effect size metric for comparison and then aggregate these studies to obtain an overall effect size estimate.

Stages in the Meta-Analysis Process

When planning and conducting a meta-analysis, it is important to follow established reporting guidelines and recommendations. The Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) is presented on a website that offers reporting guidelines and standards (<https://prisma-statement.org>) to encourage the transparent reporting of systematic reviews. Notably, the website contains a checklist of recommended information that should be reported in the title, abstract, introduction, method, results, and discussion sections of a meta-analysis manuscript. Meta-analysts are encouraged to consult with that checklist to ensure that items have been satisfied before submitting the synthesis for publication.

Understand the Literature

The first step in conducting a meta-analysis, as in any other approach to reviewing research literature, is to develop a strong understanding of the studies that have been conducted in that research area. Techniques involving meta-analysis are most effective when they are focused on a hypothesized relation involving (usually two) variables. Based on the theory and prior research, the starting point is always the simple A to B relation, whether causal or correlational. It is possible to meta-analyze interrelations involving three or more variables using an extension of the basic two-variable scenario. Returning to our earlier example, the meta-analytic hypothesis may be that a physically attractive person (independent variable) produces positive reactions such as liking (dependent variable). This cause-and-effect hypothesis connection involving two variables is an ideal candidate for meta-analysis.

A prerequisite for conducting a meta-analysis is that sufficient past research is available regarding the relation of interest. If only a few primary studies of the connection are available in the literature, the meta-analysis is premature and unlikely to produce persuasive results. Usually, a meta-analysis requires a relatively large number of existing primary studies all examine the same critical association before its utility as an integrative force can be realized. A meta-analysis requires that we have a good grasp of the literature regarding the previous studies that have been conducted. To plunge into the meta-analysis before developing a strong command of the literature is foolhardy. Meta-analysis is labor intensive. To realize halfway through the project that important primary studies have been overlooked or that the basic relation under study has not been specified correctly is too late to avoid a waste of time and resources.

Study Inclusion and Exclusion Criteria

The importance of specifying a hypothesized relation becomes apparent in this phase, which entails defining the *inclusion criteria* (or *exclusion criteria*) for primary studies to be incorporated into the meta-analysis. The issue of how broadly or narrowly to define the constructs and

operational definitions involved in the research relation impacts this choice. A meta-analytic review of the effectiveness of a specific type of psychotherapy on reducing the depression symptom of oversleeping may prove too narrow, as not enough primary studies exist. On the other hand, a synthesis of the overall effectiveness of all types of psychotherapy on reducing all forms of mental illness might be too broad, as too many primary studies exist. A synthesis involving the effect of a specific type of psychotherapy (e.g., mindfulness therapy) on depression symptoms might be manageable and a possibility for a meta-analysis. Published meta-analyses have included as few as eight (e.g., Packer, 2008) to as many as 300 studies (e.g., Rosenthal & Rubin, 1978; Smith & Glass, 1977). The number of primary studies available for synthesis varies as a function of the number of existing studies in that area and the inclusion criteria.

Choosing appropriate primary studies for inclusion or exclusion involves two processes. The first requires a tentative specification of inclusion rules; that is, which studies of the critical relation will be included in the synthesis.³ The second requires meaningful justification and rationale regarding why such inclusion criteria were applied. For example, if a meta-analyst wishes to estimate the strength of the causal connection between the chosen independent and dependent variable, then studies that made use of experimental methodology will be included in the meta-analysis. The meta-analyst might offer the rationale that correlational studies about the relation were excluded because they cannot support causal interpretations. Or, the meta-analyst may wish to include only studies published after a certain date, using the rationale that a major theoretical development did not emerge until after that point. The meta-analyst should aim to justify the inclusion criteria by providing defensible reasons. Keep in mind that overly narrow inclusion criteria may overlook important results. For example, in a meta-analysis of studies examining gender differences on perceived quality of life, Wood et al. (1989) found that men reported better quality of life in studies published after 1978, whereas women reported a better quality of life in studies published before that date. These meta-analytic results provide considerable grist for the speculative mill. If only studies published after 1978 were included and the year of each study had not been coded, this meta-analytic finding would not have emerged.

Returning to the attractiveness to liking example, we might include only studies that manipulated the independent variable of attractiveness using photographs. In terms of the dependent variable, we might include only studies that have operationally defined liking in terms of attitude scales. Or, instead, we might include studies that operationally defined liking in terms of observable behaviors such as actions, donations, and eye contact duration. The decision about the allowable breadth of the constructs must be made in terms of the goals of the meta-analysis and the number of studies available. In some circumstances, a constricted operational definition of a construct will be chosen, because this specificity is relevant to a theoretical aspect of the relation to be meta-analyzed. In other circumstances, the operational definitions will not be constricted as not enough studies are available.

There is some debate about the inclusion or exclusion of studies based on the quality of the research. Some have suggested that primary studies that do not meet some common methodological standard (e.g., no control group, no manipulation check information, poor reliability of measures, etc.) should be excluded (e.g., Greenwald & Russell, 1991; Kraemer et al., 1998). Generally, a meta-analysis should not exclude studies that satisfy inclusion criteria, even those that are methodologically suspect. Rather than including studies based on methodological purity, a more promising approach entails rating the methodological rigor of each study (e.g., weak vs. strong) and then entering this rating as a potential meta-moderator variable (meta-moderation analyses are discussed later in this chapter). This meta-moderator analysis may show, for example, that only methodologically strong, but not weak, studies exhibit the postulated relation.

This meta-analytic result could prove useful and interesting and would not be available if the weak studies were excluded.

Searching for Primary Studies

After deciding on the inclusion rules, the meta-analyst must gather every possible primary study that satisfies these criteria. This process of finding every available primary study about a research relation based on the inclusion criteria is known as an *exhaustive search*. An exhaustive search is necessary to avoid the problem known as **retrieval bias**, in which not all available primary studies that satisfied the inclusion criteria were included and synthesized in the meta-analysis. Retrieval bias may lead to overestimation or underestimation of the true underlying effect size: A synthesis based on a non-exhaustive search may not be representative of a synthesis based on all the available primary studies. An exhaustive search should be conducted by searching across research databases and applying a variety of search terms. Computerized database searches of published studies are possible by using sources such as the Social Science Citation Index, Web of Science, the Education Resources Information Center (ERIC), PsycINFO, PubMed, Sociological Abstracts, and Google Scholar. Additionally, we need to consider the appropriate search terms for the critical constructs used in the meta-analysis. Given that the same construct may be referred to by different synonyms in the literature, it is wise to employ various search terms based on the inclusion criteria. Campbell (1963), for example, listed more than 75 synonyms used in the literature to refer to the general concept that laypeople would identify as attitude or opinion. Thus, if we wished to conduct a meta-analysis on the relation between self-interest on attitudes, it would be shortsighted to exclusively use the combined search of terms “self-interest” and “attitude.” Synonyms for the self-interest construct (e.g., outcome involvement, ego involvement, hedonic relevance, vested interest) can be concurrently searched with synonyms of the attitude construct (e.g., opinion, belief, view, preference). We know the appropriate search terms by being familiar with the research literature.

Another way of using electronic databases for a literature search of eligible studies is through descendancy and ascendancy search procedures (Cooper, 2009). In **descendancy searching**, an important early and relevant primary study is identified and used as the starting point to search for all chronologically later studies (descendants) that have cited it. This primary study is considered a classic in the field and almost all subsequently published studies in that area cite that seminal study. Usually, relatively widespread consensus exists regarding such studies, and if more than one is available, all should be used to obtain additional studies in descendancy searching. Conversely, it also is practical to consider **ascendancy searching**, in which the reference list of a recently published primary study is consulted to perform a backward search to find chronologically earlier relevant studies (ancestors). These located primary studies, in turn, can be used for further backward literature searches, and the process continues until all relevant studies are found.

Some of the literature sources listed here include unpublished studies. The ERIC database, for example, contains a large store of papers presented at major scientific conferences. Unpublished studies, such as student projects, if posted online, may be found in popular Internet search engines (e.g., Google or Yahoo). Such papers are not published in a journal for various reasons, and the information would not be available if not for the Internet. Unpublished research is important to include in meta-analyses given that scientific journals prefer publishing primary studies that exhibit statistically significant results. Indeed, it is rare for journals to publish studies containing results that are primarily or all nonsignificant. Meta-analyses have demonstrated that published studies tend to show larger effects than unpublished ones. Smith and Glass (1977), for

example, examined the effect size obtained in 375 studies about the effectiveness of psychotherapies versus control groups on improvement in mental health outcomes and found that primary studies published in books and research journals had effect sizes (Cohen's *d*) that averaged between .70 and .80, whereas dissertation studies averaged .60, and other unpublished studies only .50. If the goal of the meta-analyst is to describe the universe of studies that concern a specific relation, then identifying this bias toward publishing studies with larger effect sizes would be informative about that line of research.

Computing the Effect Size for the Primary Study

If the search procedures have succeeded in producing a compilation of all primary studies relevant to the topic, the next step begins the computational phase of the meta-analysis. In this stage, the test statistics or statistical analyses reported in the primary studies are converted to the same effect size metric. An **effect size** is a standardized metric of practical significance or magnitude of the relation between two variables and serves as the main unit of analysis in a meta-analysis. The reason conversion to a common effect size metric is necessary is that different primary studies use different measurement scales and metrics. Applied to the meta-analysis example of the connection between attractiveness and liking, one primary study assessing liking may be measured with a five-point scale and another primary study may measure liking using a 100-point scale. These different metrics are not directly comparable, so we need to convert the statistical results of primary studies to the same standardized effect size metric. Calculating effect sizes is sometimes made difficult by primary studies that fail to report the necessary statistics to enable the conversion. A primary study must present results in terms of appropriate statistics (e.g., means and standard deviations, *t*-test, *F*-test, χ^2 test) that can allow conversion to an effect size.⁴ Three major families of standardized metrics are commonly used to represent effect size (Borenstein et al., 2021; Cooper et al., 2019): Cohen's *d*, odds ratio, and correlation.

Cohen's *d*. The first measure of effect size, **Cohen's *d***, is the standardized difference between two means and should be used as an indicator of effect size if the relation involves a binary variable and a quantitative variable. It is meaningful for nonexperimental studies comparing differences in means between two groups on an outcome (e.g., the extent that males and females differ on self-esteem scores). It also is appropriate for experimental designs comparing whether and to what extent two manipulated conditions of an independent variable yield different mean scores on the dependent variable. The formula for Cohen's *d* is:

$$\text{Cohen's } d = \frac{M_2 - M_1}{SD_{\text{pooled}}}$$

The difference in means is computed between the two groups – M_2 (e.g., treatment) compared to M_1 (e.g., control) – and then divided by its pooled standard deviation. The pooled standard deviation represents the overall standard deviation value across the two groups weighted by sample size (if both groups possess the same sample size, the pooled standard deviation is the mathematical average of the two standard deviations). It is necessary to divide the difference by the standard deviation as this accounts for the variability or spread of scores in each primary study. Thus, Cohen's *d* reflects the magnitude of the difference between two group means relative to the common standard deviation. Cohen's *d* can be interpreted as the difference between the two groups in terms of the number of standard deviation units or standardized units. For example, suppose that $M_2 = 115$ and $M_1 = 100$, and the $SD_{\text{pooled}} = 15$. This calculation will yield a Cohen's *d* value of $1.00 \left(\frac{115 - 100}{15} \right)$, signifying that the second group scored one standard

TABLE 20.1 Effect sizes (Cohen’s *d*) of ten primary studies

<i>Study #</i>	<i>First Author</i>	<i>Cohen’s d</i>
1	Van Damme	−.15
2	Weaver	−.05
3	Stallone	.28
4	Jovovich	.35
5	Chan	.35
6	Grier	.37
7	Schwarzenegger	.56*
8	Yeoh	.59*
9	Snipes	.65*
10	Li	.89*

**p* < .05.

deviation higher than the first group. If the result is in the hypothesized direction, with the treatment group (M_2) scoring higher than the control group (M_1), Cohen’s *d* results in a positive value. Two other possibilities may occur if the results do not support the researcher’s hypothesis. If the means for the groups are the same, the formula will yield a value of Cohen’s $d = 0$. If the results are opposite to the hypothesized direction and reveal that the control group (M_1) scored better than the experimental group (M_2), Cohen’s *d* will reveal a negative value.

In our meta-analysis of attraction to liking, Cohen’s *d* is calculated for each of the available primary studies to quantify the difference in practical effect between the attractive and unattractive conditions on liking scores. These ten primary studies are displayed in Table 20.1 and will serve as the guiding research scenario to detail the steps of a meta-analysis for the rest of the chapter. Specifically, eight primary studies each show an effect size in the direction of the meta-analytic hypothesis, but the remaining two studies obtained results in the opposite direction. Moreover, note that four of the ten studies exhibited a statistically significant result at $p < .05$.

Odds ratio. The *odds ratio* is most meaningfully interpreted as an effect size metric for studies that examine the relation between two binary variables. For instance, it may serve as an indicator of practical significance in a nonexperimental study comparing Democrats and Republicans on the outcome variable of whether they are marijuana users. Or, it may be calculated for an experimental investigation comparing treatment and control groups on a dependent variable represented by two levels. The odds ratio is less commonly reported as an effect size in social science studies but is common in biomedical studies that examine binary outcomes of success or failure (e.g., survival vs. death, healthy vs. unhealthy).

Correlation coefficient. A *correlation* is the most meaningful effect size metric for studies that examine a relation in which both variables are quantitative. The correlation coefficient is not only a type of statistical test (see Chapter 10), but it also serves as an indicator of effect size. For example, it can be used to gauge the practical magnitude in a nonexperimental study of the covariation between the level of conscientious personality and score on an intelligence test. The correlation family consists of three types: (a) the *Pearson correlation* for assessing the relation involving two quantitative variables, (b) the *point-biserial correlation* for assessing the relation involving a binary variable and a quantitative variable, and (c) the *phi correlation* for assessing the relation involving two binary variables. In fact, these three types of correlations may be combined into a meta-analysis. Given these advantageous properties, Rosenthal (1991) recommended the correlation as the preferred choice in a meta-analysis, especially if the predictor to outcome relation varies widely across primary studies in terms of whether variables were measured using a binary or quantitative metric. These three correlation indices provide information

and direction and strength of a relation and therefore permit comparison and aggregation across primary studies.

These correlation coefficients are restricted to values ranging from -1.00 to 1.00 , so they should be corrected for skewness, especially if the coefficient is near either extreme end, using *Fisher's Zr correction* (Borenstein et al., 2021). After these effect sizes based on Fisher's Z_r are aggregated across all the primary studies, the summary effect based on Fisher's Z_r should be converted back to r for interpretational purposes.

Computing the Summary Effect

After effect sizes are calculated for all primary studies, they are then aggregated. This produces a **summary effect**, an overall effect size across the primary studies that can be estimated using an unweighted, fixed-effect, or random-effects model. Applied to our attractiveness to liking meta-analysis example (Table 20.1), the weight of each effect size using each of these three weighting procedures is presented in Table 20.2. Studies one to nine each contains 100 participants, whereas Study ten contains 1,000 participants. In an **unweighted model**, the effect sizes of primary studies are mathematically averaged, without taking into consideration the sample size of each study, to calculate the summary effect. In this scenario, using the unweighted technique to combine studies, the summary effect size is Cohen's $d = .38$. In an unweighted model, the ten studies, regardless of sample size, will each contribute 10% to the summary effect. Thus, investigations of 100 or 1,000 participants will play an equal role in the estimation of the summary effect. Early meta-analyses published many years ago typically pursued a non-weighted approach. This is no longer a justifiable practice, considering that primary studies with larger samples contain more representative samples and yield more stable effect size estimates.

Use of a weighted technique is currently recommended for aggregating effect sizes in meta-analyses (Borenstein et al., 2021, Cooper, 2019; Lac, 2014). In a **fixed-effect model**, the effect sizes of primary studies containing larger sample sizes are weighted more heavily to calculate the summary effect. As shown in Table 20.2, Studies one to nine each contain 100 participants

TABLE 20.2 Computing summary effects (Cohen's d): Unweighted, fixed-effect, and random-effects models

Study #	Cohen's d	N	Percent weighted contribution to summary effect		
			Unweighted model	Fixed-effect model	Random-effects model
1	-.15	100	10.0%	5.6%	9.8%
2	-.05	100	10.0%	5.6%	9.8%
3	.28	100	10.0%	5.5%	9.8%
4	.35	100	10.0%	5.5%	9.8%
5	.35	100	10.0%	5.5%	9.8%
6	.37	100	10.0%	5.5%	9.7%
7	.56	100	10.0%	5.4%	9.7%
8	.59	100	10.0%	5.4%	9.7%
9	.65	100	10.0%	5.3%	9.7%
10	.89	1,000	10.0%	50.8%	12.4%
Summary effect (Total $N = 1,900$):			$d = .38$ ($p < .05$)	$d = .61$ ($p < .05$)	$d = .40$ ($p < .05$)

Note: For primary studies 1–9, percent contribution to summary effect varies slightly (e.g., 5.3%–5.6%) due to additional statistical adjustments (a larger absolute value of Cohen's d is slightly penalized in the weighting process) in computing the summary effect for a meta-analysis.

and therefore play a lesser role (5.3%–5.6% each) compared to Study ten with 1,000 participants and a dominating weight (50.8%) in estimating the summary effect.

Notice that Study ten has disproportionately pulled the summary effect upward to Cohen's $d = .61$. This is not an ideal situation, as an outlier study with a very large sample size relative to that of the other studies will dominate the estimation of the summary effect. A fixed-effect model assumes that a single “fixed population” of studies, addressing the same hypothesis, underlies the effect size distribution. The fixed-effect model assumes that differences in effect sizes across primary studies are attributed to a single source: sampling error (chance). In other words, the view is that all primary studies of the same research connection are conducted in the same way, and that differences in effect sizes are entirely attributed to sampling error. Essentially, the premise is that studies examining a relation are performed using exact replications, which is not consistent with the reality of how most primary research is conducted. Consequently, a major disadvantage of a fixed-effect model is that the distribution of effect sizes is expected to be homogeneous, with the meta-analysis findings only generalizable to the primary studies that were conducted in the same exact way.

In a **random-effects model**, the effect sizes of primary studies are weighted as a compromise between sample size and the number of primary studies to calculate the summary effect. It is the most appropriate estimation technique if the distribution of effect sizes is presumed to be heterogeneous. Let's examine the random-effects model of Table 20.2. Studies one to nine with smaller sample sizes will each have a smaller contribution (9.7%–9.8% each), but Study ten containing the large sample has a slightly larger contribution (12.4%) to the summary effect. Thus, a study with a large sample size would not drastically dwarf the contribution of the other studies. The summary effect calculated for the random-effects model is Cohen's $d = .40$.

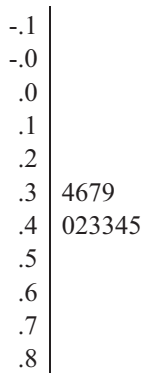
The random-effects model assumes many “random populations” of studies, addressing the same hypothesis, are represented in the distribution of effect sizes. The random-effects model assumes that differences in effect sizes across primary studies are attributed to two sources: (a) sampling error (chance) and (b) variations in operational definitions and other methodological differences across studies. In other words, the view is that primary studies regarding the same hypothesized association are conducted in different ways by different researchers. Essentially, the premise is that primary studies examining a relation are performed using conceptual replications, which is consistent with the reality of how most research is conducted. Given that the heterogeneous dispersion of effect sizes has been accounted for in the synthesis, the meta-analytic findings may be generalizable to primary studies, including future studies, that examine the research connection. This is a major advantage of the random-effects model.

In theory and in practice, random-effects models should be the preferred choice over unweighted and fixed-effect models (Borenstein et al., 2021; Hedges & Vevea, 1998; Lac, 2014). Typically, a random-effects model is an accurate reflection of how research is conducted in the literature. As an underlying connection between two conceptual variables is usually investigated in methodologically different ways by different researchers, the meta-analyst ought to account for the heterogeneous effect sizes across studies, and this counsels the use of the random-effects model. Statistically, when applying a random-effects model, it will default to a fixed-effect model if the dispersion of effect sizes is not heterogeneous and only sampling error is operating.

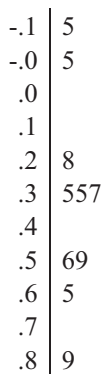
Interpreting the Meta-Analysis

Heterogeneity of Effect Sizes

Above and beyond weighting models, the interpretation of the summary effect should consider the extent of homogeneity or heterogeneity of the distribution of effect sizes across the primary



(a)



(b)

FIGURE 20.1 Stem and leaf plot of Cohen's d : homogeneous distribution of effect sizes (top) and heterogeneous distribution of effect sizes (bottom).

studies. The extent of effect size dispersion in a meta-analysis can be visually presented as a *stem and leaf plot*. A hypothetical example is illustrated in Figure 20.1a, where the effect sizes for ten studies are arrayed in a display with effect sizes organized by first and second decimal places. The values (stems) to the left of the line represent the first digit of the effect size, whereas the values (leaves) to the right of the line represent the second digit of each effect size. The number of leaves corresponds to the number of effect sizes in the meta-analysis. In this example, as ten leaves are displayed, a total of ten effect sizes are represented. This stem and leaf display shows a largely homogeneous set of effect sizes ranging from .34 to .45, with differences probably no more than expected by sampling error.

In contrast, the more dispersed the effect sizes of the primary studies over and above sampling error, the greater the heterogeneity among the individual study results. An example of a heterogeneous distribution of effect sizes involves our ten primary studies that examine the attractiveness to liking connection (Table 20.1) and is graphed using a stem and leaf plot in Figure 20.1b. Although the spread of effect size distributions differs in Figures 20.1a and 20.1b, both distributions are computed to have the same summary effect of Cohen's $d = .40$ based on a random-effects model. Obviously, the effect sizes from one primary study to another will not be identical. Some chance variation among the effect sizes is to be expected due to sampling

error. However, if the dispersion in effect sizes across the studies is above and beyond that expected from sampling error, pointing to a random-effects model, the results can be regarded as heterogeneous across the compiled primary studies.

Meta-Moderators

A **meta-moderator** is a study-level characteristic in a meta-analysis to help determine whether a set of primary studies is significantly different from another set of studies in terms of the effect size. Heterogeneity of effect sizes may provide important insights into potential study-level moderators that could help to account for these differences. A meta-analysis that reveals a high level of effect size heterogeneity provides the opportunity to investigate potential meta-moderators that are responsible for the dispersion. Meta-moderators should be tested to determine what qualifies or influences the strength of the hypothesized A to B summary effect.

We might want to qualify our meta-analytic hypothesis regarding attractiveness and liking, for example, by postulating that the relation is stronger when males compared to females are the participants or when the manipulations involve confederates rather than photographs. Including such study-level moderating variables in the meta-analysis has paid dividends when competing theorists have proposed specific characteristics and features of studies that might impinge on the fundamental relation under study. For example, Lac and Crano (2009) conducted a meta-analysis of the relation between parental monitoring and lower marijuana use in adolescents and discovered that the strength of that connection varied depending on the types of questionnaire items that primary study researchers administered to assess the construct of parental monitoring.

Meta-moderators may be classified into three varieties. *Methodological meta-moderators* include methodological characteristics, such as whether the study was conducted in a field or lab setting, whether self-report or behavioral measures were used, and whether the experimental design involved within- or between-group comparisons. *Demographic meta-moderators* include each study's average age (young, middle-aged, or older adults), socioeconomic status (wealthy or poor), and type of sample (undergraduate students or a representative sample from the population). *Theoretical meta-moderators* involve studies that used different operational definitions to define the same construct (e.g., liking was assessed using different versions of attractiveness measures) and the context under which the relation was assessed. Once potential meta-moderator variables have been proposed, each study is coded in terms of whether the characteristic is present or absent or to what degree it is present.

The effect of the study-level moderator on the hypothesized A to B relation can be tested in one of two ways depending on whether the coded meta-moderator is categorical (subgroup approach) or quantitative (meta-correlation approach). The subgroup approach involves placing all the primary studies into subgroups of studies that differ on the characteristic in question to determine whether the summary effect computed for a subgroup of studies differs significantly from another subgroup of studies (Rosenthal & Rubin, 1982). The meta-moderator variables have been added for each study in Table 20.3. The variables pertaining to gender (females only, mixed gender, or males only) and type of manipulation (photographs, confederate, or other) are categorical. To assess whether gender meta-moderated the overall summary effect, we compute the subgroup summary effect across the three studies of only female participants and compare this subgroup summary effect for the three studies involving only male participants. Analyses show that the studies containing females revealed a near zero effect size between attractiveness and perceptions of liking ($d = .03$, ns) but the association was found in male samples ($d = .60$, $p < .05$). A meta-moderator test of the statistical difference between these two subgroup summary effects reveals that the attraction to liking is significantly ($p < .05$) stronger in studies that

TABLE 20.3 Coding for three meta-moderator variables

<i>Study #</i>	<i>Cohen's d</i>	<i>Moderator 1: Gender of sample</i>	<i>Moderator 2: Type of manipulation</i>	<i>Moderator 3: Average age of sample</i>
1	-.15	Females only	Other	20
2	-.05	Females only	Other	28
3	.28	Females only	Other	33
4	.35	Mixed	Other	25
5	.35	Mixed	Photographs	24
6	.37	Mixed	Confederates	35
7	.56	Males only	Photographs	33
8	.59	Males only	Confederates	38
9	.65	Males only	Photographs	45
10	.89	Mixed	Other	55

contained male versus female participants. This result implicates gender as a meta-moderator in helping to explicate some of the heterogeneity across studies.

Next, to conduct a meta-moderation analysis for the type of experimental manipulation, we compute a subgroup summary effect for three studies using manipulations of photographs ($d = .52$, $p < .05$) and the two studies using manipulations of confederates ($d = .48$, $p < .05$). This difference in effect size is not significant and therefore indicates that type of manipulation did not account for heterogeneity in the dispersion of effect sizes.

The second method for assessing the contribution of a potential meta-moderator is used if that variable is quantitatively coded and is also known as the meta-correlation approach. In a meta-correlation, each primary study's quantitatively coded meta-moderator variable is correlated with its effect size. For an example of this meta-correlation method, consider our quantitative meta-moderator of the average age of participants in each sample, shown in Table 20.3. As older age in the sample tends to be statistically associated with a larger effect size, this meta-moderator analysis will yield a positive meta-correlation that is significant ($p < .05$). This finding would suggest that as people grow older, they may become more susceptible to liking people who are more attractive.

When the summary effect obtained in a meta-analysis is small and not statistically significant, it is important to determine that the null finding is not attributed to an effect size distribution where positive and negative effect sizes cancel each other. Suppose a researcher were to conduct a meta-analysis on prior research investigating the relation between fundamental attribution error (the tendency of people to explain others' behavior in terms of personal, rather than situational, factors) and judgment behaviors (Gilbert & Malone, 1995; Ross, 1977). If a comprehensive review of the literature were performed, the meta-analyst might find that the effect sizes varied considerably from study to study, and the summary effect suggested no systematic relation between the constructs ($r = .00$). Based on this summary effect alone, we would conclude that the fundamental attribution error does not impinge upon judgments. However, the meta-analyst might realize that almost all the positive effect sizes are found in research conducted in Western cultures (e.g., United States, Europe, Australia), whereas the effect size is near zero and sometimes reversed in non-Western ones (e.g., China, Japan, etc., see Choi et al., 1999; Fiske et al., 1998). By testing Western versus non-Western culture of (study) origin as a meta-moderator of the link from fundamental attribution error to behavior, the meta-analysis helps to understand an apparent inconsistency among studies on that topic. This meta-analytic result

should be used to inform future research regarding how cross-cultural variations in studies may qualify the strength of the fundamental attribution error on judgments.

Publication Bias

Publication bias occurs when primary studies are not published for some systematic reason, usually because results were not statistically significant or the effect size was small. But other reasons may exist for publication bias, including that research conducted by professors are more likely to be published in journals than research conducted by students. The **file drawer problem** (Bradley & Gupta, 1997; Rosenthal, 1979, 1991) reflects the tendency for unpublished studies to be tucked away in physical or electronic file drawers and unavailable for a meta-analysis, thereby resulting in a synthesis not representative of that line of research. If these “missing” unavailable studies were hypothetically found and therefore included in the meta-analysis, the overall summary effect may change. For instance, a worst-case scenario is that a meta-analysis synthesized ten published journal articles that were each statistically significant in the hypothesized direction, but 15 unpublished studies that were unavailable were each significant in the opposite direction. If the unavailable studies were incorporated into the meta-analysis, the magnitude and direction of the summary effect would be drastically different.

The extent that publication bias creates a problem for any meta-analysis may never be known fully. However, Rosenthal (1979, 1991) offers a method to estimate the extent that publication bias is a problem in a particular meta-analysis and refers to this number as the *fail-safe N*. This approach involves calculating the number of “hypothetical” nonsignificant primary studies exhibiting null results (i.e., Cohen’s d or $r = .00$ and $p = 1.00$) that would need to exist in the “file drawers” before the statistically significant summary effect obtained from the meta-analysis of included studies would change from $p < .05$ to the edge of nonsignificance at $p > .05$. The fail-safe N calculation helps to evaluate the seriousness of the threat of publication bias to the conclusions drawn from the meta-analysis. If the number of additional unpublished studies that would need to exist is very large before the observed summary effect is rendered nonsignificant, then it is not likely that studies in the “file drawers” would compromise the conclusions. But if only a few additional unpublished studies with null results would be needed to reduce our obtained summary effect to the edge of statistical nonsignificance, the researcher would regard the file drawer problem as a potential threat to the interpretation of the computed meta-analysis.

Another method of evaluating potential publication bias effects is by inspecting the stem and leaf plot to determine the extent that the effect sizes are symmetrical and preferably normally distributed. For example, Figure 20.1a and 20.1b show two stem and leaf plots. Both reveal approximately a symmetrical distribution of effect sizes, suggesting that publication bias is unlikely in these two meta-analyses. The symmetry in these two examples may not be evident due to the inclusion of only ten primary studies, but a stem and leaf plot consisting of many more studies would provide more conclusive results about the extent of publication bias. An asymmetrical or skewed distribution of effect sizes would signify publication bias.

The statistical reason why a symmetrical normal distribution of effect sizes signifies no publication bias is explained as follows. Most primary studies in an area of research will generate effect sizes that are relatively close in value to the summary effect size, but effect sizes closer to the extreme low or extreme high of the effect size range will be less common. Moreover, about half the primary studies are expected to yield a smaller effect and the other half a larger effect size compared to the summary effect. A disadvantage of the stem and leaf plot in helping to judge publication bias is that it considers only effect sizes and ignores the sample sizes of the primary studies. The consideration of sample size is important when evaluating publication

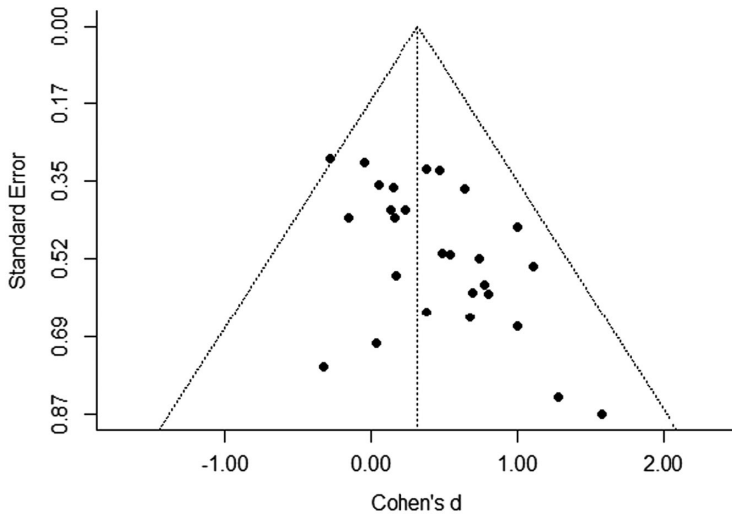


FIGURE 20.2 Funnel plot based on 28 available effect sizes.

bias, because primary studies containing smaller sample sizes are less likely to attain statistical significance and therefore less likely to be published. The most reported advanced publication bias test in meta-analysis is called the funnel plot (Egger et al., 1997). In a **funnel plot**, the primary studies are graphed in a scatter plot to inspect the distributional pattern of effect sizes and to evaluate the extent of publication bias. An example of a funnel plot is shown in Figure 20.2. The X-axis represents the effect size and the Y-axis represents the standard (sampling) error for each study (standard error is inversely associated with sample size). A primary study possessing a smaller standard error denotes a larger sample size, whereas a primary study possessing a larger standard error denotes a smaller sample size. In published meta-analyses, it is established practice for funnel plots to display the standard error (rather than the sample size) on the Y-axis, as the standard error is more directly connected to the p-value of a primary study.

The vertical dotted line in the middle of the funnel plot reveals the summary effect across all the primary studies. The sides of the pyramid, or upside-down funnel, denote the lower and upper bound for the 95% confidence intervals around the summary effect size. Notice that a smaller standard error (larger sample size) will yield a narrower confidence interval – consistent with the statistical notion that primary studies based on larger sample sizes will yield greater confidence or precision in estimating the summary effect size. A primary study containing an effect size that is inside the pyramid indicates that the effect is not significantly different from the summary effect; that is, the difference between a primary study effect and the summary effect is attributed to sampling error. As such, the funnel plot can be used to visually inspect whether the distribution of effect sizes is homogeneous or heterogeneous. If all the scatter points fall inside the pyramid, the effect sizes are homogeneous. However, the more scatter points that fall outside the pyramid, the more likely the effect sizes are heterogeneous, and consequently a random-effects model is warranted.

Finally, publication bias is determined by inspecting the extent of symmetry of the scatter plot points around the summary effect size in the funnel plot. If the number of scatter points is approximately symmetrical on the left and right sides, then little or no publication bias is demonstrated. Instead, if the number of points on the left and right sides is asymmetrical, then this suggests publication bias. For example, if scatter points are more heavily distributed toward

the left side, this suggests that “missing” studies to the right side would be needed to furnish a symmetrical funnel plot. If those missing studies were hypothetically included on the right side, then the summary effect size would be larger. Conversely, if points are more heavily distributed toward the right side, this suggests that “missing” studies to the left side would be needed to yield a symmetrical funnel plot. Hence, if those missing studies were hypothetically found and incorporated into the meta-analysis, the summary effect size would be smaller or toward a negative effect size.

Figure 20.2 shows the funnel plot of effect sizes for 28 primary studies. The vertical dotted line in the middle shows the summary effect of Cohen’s $d = .35$, $p < .05$ (this example was calculated using a random-effects model). All except for one effect size is found to be inside the pyramid (95% confidence intervals around the summary effect), indicating that the effect size distribution is primarily homogeneous. The figure also exhibits asymmetry, as most of the effect sizes are located on the right side of the summary effect, thereby signifying publication bias.

The funnel plot does not provide information regarding where the missing studies ought to be in the figure to produce a symmetrical distribution. Thus, meta-analysts are only able to eyeball the funnel plot to get a rough estimation of the extent of publication bias. To address this concern, the *trim and fill analysis* (Duval & Tweedie, 2000) was developed as a statistical supplement to the funnel plot. The trim and fill analysis mathematically imputes where the “missing” studies ought to be inserted to make the funnel plot symmetrical around the summary effect.⁵

Interpreting Summary Effects

In a primary study, the statistical power to detect a significant result is in part a function of the number of participants. In a meta-analysis, owing to pooling the number of studies and number of participants across studies, greater statistical power to attain statistical significance is afforded compared to a single primary study. Thus, a meta-analysis is ideal for synthesizing areas of primary research typified by small samples to attain greater statistical power. For example, a meta-analysis evaluated the mathematical abilities of human infants in terms of how many seconds they visually stared at the addition and subtraction of physical objects such as toys and puppets (Christodoulou et al., 2017). In this area of research, infants have been theorized to stare longer (dependent variable) at a mathematically impossible (e.g., $2 + 1 = 4$) versus possible (e.g., $2 + 1 = 3$) display of physical objects (independent variable). As newborns are extremely difficult to recruit for research, these primary studies ranged only in sample size from seven to 46 infants. Consequently, many of these investigations possessed insufficient statistical power and therefore were not statistically significant. After the meta-analysis pooled across these studies involving small sample sizes to yield a total sample of 550 infants, the summary effect size emerged as statistically significant and supported the position that infants possess some rudimentary mathematical abilities.

Effect sizes can be interpreted as small, medium, or large in terms of the practical strength of the relation between two variables. Cohen (1988) and Rosnow and Rosenthal (1996) provide recommended values and computational advice for interpreting the effect size magnitude for primary studies and summary effects. These recommendations are displayed in Table 20.4. This table provides a convenient shorthand guideline for describing the practical strength of a relation. These recommendations represent rough guidelines of strength and should not be rigidly interpreted. A major reason is that effect sizes are not as meaningful when comparisons are made across widely different domains of research. A small effect for a drug treatment that reduces the risk of a debilitating disease or saves lives is more practically desirable than a large effect obtained in a treatment designed to improve people’s emotional bond with their pet cats.

TABLE 20.4 Effect size interpretational guidelines

<i>Interpretation of practical effect</i>	<i>Cohen's d</i>	<i>Odds ratio</i>	<i>Correlation</i>
Small	.20	1.50	.10
Medium	.50	2.50	.30
Large	.80	4.00	.50

The interpretation of effect sizes is most meaningful when the meta-analyst considers both the research context (What are the consequences of the findings?) and knowledge contribution (What insights are afforded to help understand this line of research?) of the meta-analysis results (Ellis, 2010).

Apparently small effect sizes discovered in a meta-analysis should not be dismissed lightly. Rosenthal and Rubin (1982) proposed a visual method of interpreting effect size in terms of differences in positive or negative outcomes that are found between the treatment groups that form the basis of the meta-analysis. Applying the *binomial effect size display* (BESD) to interpret the summary effect provides graphic evidence of the practical importance of a relation between two variables.

Suppose, for example, that a meta-analyst is interested in the hypothesis that manipulated message source (in-group or out-group) will have an impact on participants' ratings of persuasive effectiveness. The researcher gathers studies that investigate this relation. The results suggest that in-group sources tend to be more persuasive than out-group sources, but the summary effect size is small ($r = .10$), according to Cohen's interpretational guidelines. Rather than becoming discouraged, the meta-analyst turns to the BESD and calculates the difference between low and high persuasive as a function of the manipulated sources, as shown in Figure 20.3. The BESD requires the assumption that a total of 200 participants is represented, even if the meta-analysis involves a much larger sample size. This reveals that when the source of persuasion is in-group (rather than out-group), participants are more likely to be high (rather than low) in being persuaded. Specifically, ten more participants (corresponding to $r = .10$) in the in-group compared to the out-group are high (vs. low) in being persuaded. A marketing executive might find this result to be a very important argument for targeting messages to potential consumers of a product. Although the effect size is considered small, it is well worth considering in practical terms that this strategy might yield millions of dollars in product sales.

	In-Group	Out-Group
High Persuasiveness	n = 55	n = 45
Low Persuasiveness	n = 45	n = 55
	n = 100	n = 100
	N = 200	

FIGURE 20.3 Binomial effect size display: summary effect $r = .10$.

Vote Counting versus Meta-Analysis

Conclusions of the research literature drawn from the meta-analysis may yield different conclusions from the more traditional narrative reviews. Prior to the advent of meta-analysis and its techniques, attempts were made to make traditional narrative reviews more quantitative. **Vote counting** is a method of literature review that involves counting and comparing the number of primary studies that are statistically significant or not (Cooper et al., 2019). For example, a

vote counter may identify ten studies that investigate a particular hypothesis and then determine whether each one is statistically significant at $p < .05$. If most of the studies are statistically significant (e.g., six studies significant, but four studies nonsignificant), then the vote counter concludes that the connection is supported. If vote counting reveals that most of the studies are nonsignificant (e.g., four studies significant, but six studies nonsignificant), then the meta-analyst concludes that the relation does not exist. A major problem with vote counting is that the decision of whether a research link is tenable is usually made based on such borderline votes.

The direct comparison of p -values of primary studies is generally not wise, but the vote counting approach is completely dependent on such black and white cut-offs. The problem with vote counting is twofold. First, p -values between different studies are not directly comparable, given that larger sample sizes are more likely to attain statistical significance. Second, this approach loses considerable information as the magnitude of the effect size is completely ignored. Vote counting relies strictly on the dichotomous distinction of study significance or nonsignificance, whereas the overall effect obtained in a meta-analysis offers more nuanced results about the magnitude of the summary effect and its statistical significance (Cooper & Rosenthal, 1980).

To illustrate the differences in conclusions that might be drawn from meta-analytic versus vote counting, consider again the data in Table 20.1. Only four of ten studies reported statistically significant effects, although most of the studies yielded outcomes in the same positive direction. In this scenario, a vote counting analysis would have led us to conclude erroneously that no connection existed between attractiveness and liking. Conversely, the results of the meta-analysis showed a statistically significant summary effect, with a value that indicates the magnitude of the relation. The difference in conclusion stems from accumulating the sample sizes from all the studies and the use of effect sizes in meta-analysis to estimate the strength of effects. In a meta-analysis, even studies with small or nonsignificant effect sizes add information about the magnitude and direction of the summary effect.

Conclusion

Meta-analysis has much to offer in quantitatively consolidating and synthesizing research on a given topic. It promises to facilitate the growth and development of the research enterprise and therefore should become a standard practice when seeking to synthesize a line of research. This is not to suggest that meta-analysis should supplant primary data collection or even narrative reviews. Meta-analysis obviously cannot supplant primary research because original data are the fundamental inputs for any meta-analysis. Primary research produces data for inclusion in future meta-analyses, whereas the conclusions drawn from a meta-analysis may, in turn, foster and inform subsequent primary research.

Clearly, meta-analysis should and will play a larger role in the research enterprise. Narrative reviews complement and enrich the interpretation of meta-analyses. The two approaches of reviewing the literature are best used in combination. Particularly where there are substantial variations in effects obtained across studies examining the same hypothesis, careful consideration of meta-moderators examining demographic, methodological, and theoretical variations among the studies included in a meta-analysis is essential for drawing meaningful conclusions.

Notes

- 1 Of course, it is not fair to blame the narrative method for reviewers' failures to perform complete literature searches. However, the tradition of the narrative method does not necessarily call for complete surveying, which often produces a biased review.

- 2 Stigler (1986) finds evidence of rudimentary quantitative synthesis in the 1800s; Thorndike (1933) provides a somewhat more contemporary example.
- 3 The specification is tentative because as the meta-analysis progresses, the rules may be changed to account better for the available data.
- 4 Authoritative sources for conversion formulas are authored by Borenstein et al. (2021) and Cooper et al. (2019).
- 5 The trim and fill analysis can impute the hypothetical summary effect if the hypothetical missing studies were incorporated into the synthesis. If publication bias no longer exists and the synthesis involved 35 effect sizes (28 available and seven imputed), this will produce a hypothetical summary effect of Cohen's $d = .22$, $p < .05$. Notice that this imputed summary effect is smaller than the original summary effect (Cohen's $d = .35$, $p < .05$), as the now included missing studies on the left side of the funnel plot are pulling down the magnitude of the summary effect.

References

- Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2021). *Introduction to meta-analysis*. Hoboken, NJ: John Wiley & Sons.
- Bradley, M. T., & Gupta, R. D. (1997). Estimating the effect of the file drawer problem in meta-analysis. *Perceptual and Motor Skills*, 85, 719–722.
- Campbell, D. T. (1963). Social attitudes and other acquired behavioral dispositions. In S. Koch (Ed.), *Psychology: A study of a science* (Vol. 6). *Investigations of man as socius* (pp. 94–172). New York: McGraw-Hill.
- Carey, B. (2015). Many psychology findings not as strong as claimed, study says. *New York Times*, August 27.
- Choi, I., Nisbett, R. E., & Norenzayan, A. (1999). Causal attribution across cultures: Variation and universality. *Psychological Bulletin*, 125, 47–63.
- Christodoulou, J., Lac, A., & Moore, D. S. (2017). Babies and math: A meta-analysis of infants' simple arithmetic competence. *Developmental Psychology*, 53, 1405–1417.
- Clark, R. D. (1971). *Einstein: The life and times*. New York: World Publishing.
- Cohen, J. A. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.) (2019). *The handbook of research synthesis and meta-analysis*. New York: Russell Sage Foundation.
- Cooper, H. M. (1979). Statistically combining independent studies: A meta-analysis of sex differences in conformity research. *Journal of Personality and Social Psychology*, 37, 131–146.
- Cooper, H. M. (2009). *Research synthesis and meta-analysis: A step-by-step approach* (4th ed.). Los Angeles, CA: Sage Publications.
- Cooper, H. M., & Rosenthal, R. (1980). Statistical versus traditional procedures for summarizing research findings. *Psychological Bulletin*, 87, 442–449.
- Crandall, C. S., & Sherman, J. W. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology*, 66, 93–99.
- Duval, S., & Tweedie, R. (2000). A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56, 455–463.
- Edlund, J., Cuccolo, K., Irgens, M., Wagge, J., & Zlovovich, M. (2022). Saving science through replication studies. *Perspectives on Psychological Science*, 17, 216–225.
- Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315, 629–634.
- Ellis, P. D. (2010). The essential guide to effect sizes: *Statistical power, meta-analysis, and the interpretation of research results*. Cambridge: Cambridge University Press.
- Fiedler, K., & Prager, J. (2018). The regression trap and other pitfalls of replication science – Illustrated by the report of the open science collaboration. *Basic and Applied Social Psychology*, 40, 115–124.
- Fiske, A. P., Kitayama, S., Markus, H. R., & Nisbett, R. E. (1998). The cultural matrix of social psychology. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (Vol. 2, pp. 915–971). New York: Oxford Press.
- Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin*, 117, 21–38.
- Greenwald, S., & Russell, R. I. (1991). Assessing rationales for inclusiveness in meta-analytic samples. *Psychotherapy Research*, 1, 17–24.
- Hedges, L. V. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, 4, 486–504.

- Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, 3, 486–504.
- Klein, R., et al. (2014). Investigating variation in replicability: A “Many Labs” replication project. *Social Psychology*, 45, 142–152.
- Klein, R., et al. (2018). Many Labs 2: Investigating variability in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1, 443–490.
- Kraemer, H. C., Gardner, C., Brooks, J. O. III, & Yesavage, J. A. (1998). Advantages of excluding underpowered studies in meta-analysis: Inclusionist versus exclusionist viewpoints. *Psychological Methods*, 3, 23–31.
- Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press. (1st ed., 1962).
- Lac, A. (2014). A primer for using meta-analysis to consolidate research. *Substance Use and Misuse*, 49, 1064–1068.
- Lac, A., & Crano, W. D. (2009). Monitoring matters: Meta-analytic review reveals the reliable linkage of parental monitoring with adolescent marijuana use. *Perspectives on Psychological Science*, 4, 578–586.
- LeBel, E. P., Campbell, L., & Loving, T. J. (2017). Benefits of open and high-powered research outweigh costs. *Journal of Personality and Social Psychology*, 113, 230–243.
- Miller, J., & Ulrich, R. (2022). Optimizing research output: How can psychological methods be improved? *Annual Review of Psychology*, 73, 691–718.
- Nosek, B., et al. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, 73, 719–148.
- Open Science Collaborative (2015). Estimating the reproducibility of psychological science. *Science*, 349 (6251), acc 4716
- Packer, D. J. (2008). Identifying systematic disobedience in Milgram’s obedience experiments. *Perspectives on Psychological Science*, 3, 301–304.
- Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin*, 85, 185–193.
- Rosenthal, R. (1979). The “file drawer problem” and tolerance for null results. *Psychological Bulletin*, 86, 638–641.
- Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Newbury Park, CA: Sage.
- Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences*, 1, 377–415.
- Rosenthal, R., & Rubin, D. B. (1979). Comparing significance levels of independent studies. *Psychological Bulletin*, 86(5), 1165–1168.
- Rosenthal, R., & Rubin, D. B. (1982). Comparing effect sizes of independent studies. *Psychological Bulletin*, 92, 500–504.
- Rosnow, R. L., & Rosenthal, R. (1996). Computing contrasts, effect sizes, and counternulls on other people’s published data: General procedures for research consumers. *Psychological Methods*, 1, 331–340.
- Ross, L. (1977). The intuitive scientist and his shortcomings. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol 10, pp. 174–221). New York: Academic Press.
- Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, 32, 752–760.
- Stanley, D. J., & Spence, J. R. (2014). Expectations for replications: Are yours realistic? *Perspectives on Psychological Science*, 9, 305–318
- Stigler, S. M. (1986). *History of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Harvard University Press.
- Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, 9, 59–71.
- Thorndike, R. L. (1933). The effect of the interval between test and retest on the constancy of the IQ. *Journal of Educational Psychology*, 25, 543–549.
- Van Bavel, J., Mende-Siedlecki, P., Brady, W., & Reinero, D. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences*, 113, 6454–6459.
- Wood, W., Rhodes, N., & Whelan, M. (1989). Sex differences in positive well-being: A consideration of emotional style and marital status. *Psychological Bulletin*, 106, 249–264.

SOCIAL RESPONSIBILITY AND ETHICS IN SOCIAL RESEARCH

The social sciences share with all scientific endeavors the need to balance scientific zeal with other values that derive from the social context in which all scientific work takes place. To some extent, the scientific ideal of objectivity inevitably conflicts with humanistic values, and all scientists must at some time come to grips with this conflict. The issue, however, is particularly acute for social scientists because the focus of their research is the behavior of other human beings, and thus not only the goals of research but the very process of data collection is subject to value conflicts. In this chapter we consider first those ethical issues that are related specifically to research with human respondents; we then focus on the more general ethical issues of concern to all scientists in their choice of research problem and strategy of data reporting.

Ethics of Research Practices

Because the subject matter of the social scientist is human behavior, and the processes responsible for behavior, it is inevitable that scientific interests might conflict with values placed on the rights of individuals to privacy and self-determination. The guidelines for behavioral and social research set by research societies, such as the American Psychological Association's Committee on Ethical Standards (2002; see also Knapp, 2011), and the President's Panel on Privacy and Behavioral Research (Surgeon General's Directive, 1967), stress the necessity for **informed consent**, a process in which potential participants are informed about what participation in the study will involve, so that they can make an informed decision whether or not to engage in the research. Doing so emphasizes that participation should be voluntary. It is recognized that many phenomena could not be researched at all if this ideal were fully satisfied, and that the rights of participants must be weighed against the potential importance of the research problem. Resnick and Schwartz (1973), for example, demonstrated in a verbal conditioning experiment that providing a complete description of methods and hypotheses to participants eliminated the phenomenon they were attempting to investigate.

In studies where full disclosure about the hypothesis and purpose of the study cannot be provided to participants, the panel report recommends that "consent be based on trust in the qualified investigator and the integrity of the research institution." Thus, the ethical code does not provide absolute standards that relieve the scientist of important value judgments. Rather,

judgments as to the relative importance of research programs and researchers' responsibilities for the welfare of their participants are the fundamental bases of research ethics.

Deception in the Laboratory

When research is conducted in the laboratory, there is seldom any doubt concerning participants' knowledge that they are the subjects of scientific investigation. The extent to which participation is fully voluntary is debatable, given the social and institutional pressures to take part in research that are sometimes involved. But generally, due to the artificial setup, participants in laboratory studies at least *know* that they are taking part in a research study. Beyond that, the information provided to participants in laboratory investigations is usually minimal to foster research purposes, and sometimes is intentionally misleading by design.

As we emphasized in Chapter 6, to satisfy the goal of measuring reactions that are as natural as possible, the methodological strategy of most laboratory research is directed toward motivating participants to behave spontaneously and therefore not self-consciously while involved in the conditions of the study. To this end, research procedures might be presented and justified in the guise of an elaborate "cover story" designed to manage participants' perceptions of the research setting and (sometimes) conceal the true purpose of the study. Thus, the researcher often goes beyond merely withholding information from the participant to deliberate misrepresentation of the details of his or her participation. The importance of this form of researcher control in experimental designs has been discussed in previous chapters of this book. Yet, such deception is undeniably in violation of the values of interpersonal trust and respect.

To what extent this violation is justified in the service of scientific goals and the potential advancement of human welfare is a matter of considerable controversy. Some critics have argued that deception is never justified and should not be permitted in the interests of social research (e.g., Ortmann & Hertwig, 1997). Most researchers take a more moderate view, recognizing that there is an inevitable trade-off between the values of complete honesty in informed consent and the potential value of what can be learned from the research itself. Just as the "white lie" is regarded as acceptable when used in the service of good manners, a minimal amount of deception may be tolerated in the service of obtaining trustworthy research data. However, there is some debate over whether behavioral scientists have exceeded this acceptable threshold of deception in their research.

For some years the practice of deception in experiments was accepted with equanimity by most experimenters. However, an article by Kelman (1967) reflected a growing concern with the widespread, and apparently unquestioned, use of deception in research in social research. Kelman's article called into question this practice on both ethical and practical grounds. Ethically, he argued, any deception violated implicit norms of respect in the interpersonal relationship that forms between experimenter and research participant. In addition, the practice might have serious methodological implications as participants become less naïve and widespread suspiciousness begins to influence the outcomes of all human research. To offset these problems, Kelman recommended that social scientists (a) reduce unnecessary use of deception, (b) explore ways of counteracting or minimizing the negative consequences stemming from deception, and (c) use alternative methods, such as role-playing or simulation techniques, which substitute active participation for deception, with participants being fully informed of the purpose of the study.

Experimentation with these alternative methodologies has been attempted, but the results are mixed, at best (see Chapter 6), and it remains uncertain whether the results of a role-playing simulation can be interpreted in the same way as those obtained under the real experimental

conditions. Thus, the consensus in the research community is that some level of deception sometimes is necessary to create realistic conditions for testing research hypotheses, but that such deception needs to be justified in light of the purpose and potential importance of the research question being studied. As Rosenthal (1994) puts it,

The behavioral researcher whose study may have a good chance of reducing violence or racism or sexism, but who refuses to do the study simply because it involves deception, has not solved an ethical problem but only traded one for another.

(p. 129)

The various disciplines within the social sciences vary in the degree of tolerance for deception that is normative, with most fields of psychology accepting the practice of deception in experimental research under certain circumstances whereas the field of behavioral economics generally rejects the need for deception in experimental studies under any circumstances (but see Charness et al., 2022).

Deception and participant well-being. Apart from the issue of the relationship between experimenter and participant, the ethical acceptability of some deception experiments has been challenged on the grounds that the situations set up by the experimenter sometimes place research participants in a position of psychological distress or other discomfort. This potentially violates the other major canon of ethical research – to “do no harm” to those who participate. The extent to which such potential distress to participants is tolerable in the name of scientific research also is a matter of debate. As usual, the extremes are fairly well established – the risk of permanent physical or psychological harm without informed consent by the human participant is never permissible. However, consensus as to the acceptability of temporary or reversible psychological distress is more difficult to achieve. The Asch line judgment studies (1956), for example, which provided the paradigm for much research on social conformity, clearly placed the naïve participant in a position of psychological distress in weighing the evidence of one’s own perceptions against the peer pressure of judgments made by presumably sincere fellow students (who were actually confederates). Most researchers seem to agree that the potential contribution of this line of research and the transitory nature of the psychological discomfort involved justified its undertaking.

The studies of destructive obedience conducted by Milgram (1974), on the other hand, aroused considerably more variability in reaction. The design of Milgram’s research on blind obedience to an authority figure, which has been mentioned previously in this textbook, involved deceiving participants into believing that they were administering a series of escalating electrical shocks to a fellow participant (the confederate was not really receiving any shocks), while being pressured by the experimenter to continue the procedure. In initial reports of the results of this research, Milgram himself gave detailed illustrations of evidence of psychological stress on the part of those participants who continued in the experiment. Of one such participant he observed “...within 20 minutes...was reduced to a twitching, stuttering wreck, who was rapidly approaching a point of nervous collapse” (Milgram, 1963, p. 377).

As a consequence of these rather dramatic depictions of participant distress, the publication of Milgram’s initial experiments was met with much critical reaction (e.g., Baumrind, 1964). Milgram’s (1964) response to this criticism emphasized the significance of the research, particularly regarding the unexpectedly high percentage of people who were willing to obey. He noted that great care had been exercised in post-experimental information sessions to assure that participants suffered no long-term psychological damage. Milgram also suggested that the vehemence of the critical response to his research may have been largely a function of the nature of

his findings: Many participants were willing to shock the innocent victim (i.e., the confederate) all the way to the maximum voltage level. If more participants had behaved in a humane and independent fashion and refused to administer shocks at high voltage levels, the same research procedures might not have come under such intense attack. If so, it perhaps was not so much the use of deception that was being criticized – the cover story disguised it as a learning study, with a shock to be delivered every time the victim made a mistake in word memorization – but the way participants reacted to the situation. Yet those very results – unexpected as they were – are what made the experiments so valuable to social science.

Debriefing. Debate over the ethical acceptability of Milgram's experiments has revealed that whereas social scientists disagree about the frequency and extent that the use of deception should be tolerated, much research involving deception has been deemed permissible. Where deception is viewed as necessary to the exploration of some research questions, attention turns to the best method of counteracting the negative implications of its use. For this purpose, considerable emphasis has been given to the importance of the **debriefing**, a session following participation in a study, in which the participants are informed of the true purpose of the research. Kelman (1967) placed enough value on such debriefing procedures to regard their inclusion in the experimental design as an "inviolable" rule. Milgram's (1964) justification of his research procedures relied heavily on his use of extensive debriefing sessions during which participants were reassured about the nature of their responses and encouraged to express their reactions in what was essentially a psychotherapeutic setting.

When such attention is devoted to the content and handling of debriefing sessions, this may serve not only to "undeceive" the participants (and thereby relieve the researcher's pangs of conscience) but also to enrich the participant's own experience in terms of understanding human behavior and the nature of social research. However, we must caution against considering the debriefing session as a panacea for all ethical aches and pains of social research. Some research suggests that when used routinely or handled perfunctorily, debriefing procedures can cause more harm than good (Holmes, 1976a, 1976b; Ross et al., 1975; Walster, Berscheid et al., 1967).

Many researchers warn that routine debriefing may produce widespread suspicion and potential dishonesty among populations frequently tapped for participation in psychological research. Some research has been directed toward this issue, the results of which are interesting, though not entirely consistent. Brock and Becker (1966) exposed participants to one experiment using deception followed by no debriefing, partial debriefing, or complete debriefing, and then tested the same participants again in a subsequent deception experiment. No differential reactions to the second experiment were obtained except when a strong deception clue was made available through a common element in the two experiments. Results revealed that only the completely debriefed participants, as classified in the first experiment, reacted with suspicion to the second experiment. On the other hand, Silverman et al. (1970) compared students who participated in an experiment with deception to others who participated in an experiment without deception. Subsequently, significant differences in responses to a series of psychological tests emerged between these two conditions. Examination of these differences indicated that the previously deceived participants (possibly to correct for a deceived self-image) responded with more favorable self-presentation biases than did the nondeceived participants.

The preceding two studies involved students who had participated in only one experiment prior to testing. Other studies have investigated the effects of multiple participations on attitudes and behavior related to behavioral research. Holmes (1967) found that the more prior research studies that participants had engaged in, the more favorable were their attitudes toward psychological experiments, and the more they intended to take part in future research. Holmes's study, however, did not consider the extent to which his participants had experienced deception in

the studies in which they had served. A more direct test of the effects of frequent deception was provided by Smith and Richardson (1983), who tested the attitudes of more than 400 research participants who had served in a variety of experiments over the course of an academic quarter. They found that participants who had been in deception experiments evaluated their research experience more positively than those who had not been deceived, and that debriefing appeared to eliminate the negative feelings of those who felt they had been harmed by deceptive practices. Those who had participated in experiments using deception reported that they received better debriefing, enjoyed their experiences more, received greater educational benefits from the research, and were more satisfied with the research program than those who had not been involved in experiments with deception.¹

We speculate that some of these variations might be a function of differences between the types of experiments that made use of deception and those that do not. Often, deception-based studies are inherently interesting – they involve complex decision-making, arguing against a strong counter-attitudinal communication, making judgments about people and social events, and so on. Those that do not involve deception might require participants to memorize long lists of nonsense syllables, to respond as rapidly as possible to randomly presented visual stimuli, to decide which two of three tones are most similar, etc. These types of studies often are ethically pristine, but they might not be very enjoyable for participants. With effective debriefing, the negative aspects of deception studies might be offset, leading to more positive evaluation of the research experience on the part of participants. This requires, however, that the debriefing be conducted in such a way, and with enough thought and effort, that participants leave feeling that they have learned something by the experience and that their time has not been wasted.

Effective debriefing is also important in gaining participants' commitment to the purposes of the research and their agreement not to tell potential future participants about the deception employed. As a result of communication among members of a research participant pool, the effects of debriefing-generated suspicion may extend beyond those who have participated in deception experiments. Weubben (1965) provided some indication of the extent of such inter-subject "contamination" by finding that of 113 participants who had agreed to secrecy following an experimental debriefing session, 72 revealed the nature of the research to other potential participants. Reviews of deception research (Striker, 1967) indicate that there is inadequate assessment of such prior knowledge or suspicion on the part of participants and that the implications for the internal validity of experiments are still not well understood. Because of this, badly done debriefings – those that fail to gain participants' confidence, trust, and cooperation – can potentially harm the scientific enterprise.

Apart from these methodological considerations, debriefing may not always be effective in counteracting the lingering negative effects of certain kinds of deception manipulations after participation in the study. Walster et al. (1967) reported that for highly concerned participants, even lengthy debriefings were not successful in removing reactions to false information feedback regarding personal adequacy. Ross et al. (1975), too, found that the effects of deception sometimes were exceptionally difficult to offset.

Particularly when the experimental manipulation has involved providing false information of some kind about the participant's personality, competence, or self-esteem, an extensive form of debriefing known as *process debriefing* (Ross et al., 1975) may be required. Process debriefing entails discussing with participants how the deception may have temporarily influenced their own self-perceptions and the processes by which these effects might occur. Previous research has shown that this type of systematic debriefing is more successful in eliminating any lingering effects of the deception than a standard debriefing in which participants are simply told that the information was untrue but without discussion of process (Ross et al., 1975).

Another technique for enhancing the effectiveness of debriefing is to show participants all the experimental conditions so they can see how the information they had received was just one version of the experiment to which they had randomly been assigned. For instance, if participants during the experiment had been given false feedback that their test performance was negative, they could be shown the materials for the positive feedback condition (and perhaps asked to imagine how they would have reacted to that feedback), to make more apparent that the nature of the feedback they had received was due solely to chance.

Ethical Issues in Nonlaboratory Research

Although much of the debate about the ethical implications of deception focuses on laboratory experimentation, research conducted outside the laboratory often raises a host of other ethical issues and concerns. In addition to issues related to consent to participate, researchers also must consider issues of privacy and confidentiality when research data are collected in field settings. Because a major advantage of field research, from a scientific standpoint, is the possibility of obtaining samples of behavior under naturally occurring conditions, it frequently is advantageous to conduct such studies under conditions in which the nature of the research is disguised. Thus, the participants may not only be deceived about the purpose of the research but may even be unaware that they are the subject of research in the first place. The use of “unobtrusive” measures (Page, 2000; Webb et al., 1981) highlights this strategy, but even more traditional methods of data collection, such as the interview or questionnaire, are frequently presented in such a way as to disguise their true purpose.

Some scientists regard the practice of concealed observation or response elicitation as acceptable as long as it is limited to essentially “public” behaviors or settings normally open to public observation. Campbell (1969), for instance, provided a review of settings and behaviors for which disguised research strategies have been employed, including studies ranging from pedestrian violations of traffic signals (Lefkowitz et al., 1955), to mailing of “lost letters” (Milgram, 1969), interpersonal interaction in employment offices (Rokeach & Mezei, 1966), arranged encounters between strangers in public streets (Feldman, 1968), fund collection (Bryan & Test, 1967), and door-to-door solicitation (Freedman & Fraser, 1966). All of these involve behaviors that Campbell regarded as falling within the public domain and thus not requiring permission from participants nor subsequent debriefing. However, there remains the question of subjective definitions of what constitute “public” behaviors, particularly in settings where social norms lead to the expectation of anonymity in public places. Some social scientists (e.g., Miller, 1966) regard any form of unaware participation in research as an intolerable invasion of the individual’s right to privacy. Although some settings may readily allow for interpersonal observation by the researcher, if individuals in these settings do not normally expect to be observed (or, rather, expect not to be), the issue of privacy remains. A case in point is provided in the research of Middlemist et al. (1976), who studied the effects of “spatial invasion” in a men’s bathroom by secretly observing participants’ behavior at a urinal.

As field studies using an experimental design involve some intervention on the part of the researcher in determining the stimulus conditions to which the unaware participants are exposed, ethical considerations about hidden observation are further complicated by concern over such manipulations. Examples of experimentation in field settings include systematic manipulation of the content of applicant resumes sent to prospective employers (Schwartz & Skolnick, 1962), differential behavior on the part of salesmen toward customers (Brock, 1965) or customers toward salesmen (Jung, 1959; Schaps, 1972), varied content of requests for a dime from passing strangers (Latané, 1970), and the apparent condition of the victim of a feigned collapse in a New

York subway train (Piliavin et al., 1969). To varying degrees, these all fall within a normal range of human experiences in public places, the only difference being their systematic manipulation by the researcher. Yet, collecting data about people's behaviors in these situations clearly violates the spirit of informed consent, especially when researchers decide it is best not to debrief those who have been observed even after the fact.

Privacy on the Internet. The advent of the Internet as a venue for social research creates a new wrinkle in the continuing issue of what constitutes "private" behavior or invasion of privacy. In addition to archival records of communications that are exchanged through various Web-based interest groups and online forums, researchers are more and more often becoming involved as "participatory observers" in such message groups (Walther et al., 2018). Currently, the prevailing philosophy is that preexisting content on the Internet constitutes public domain, and therefore can be observed and recorded without obtaining consent, although there is an expectation that information about identity of the senders should be protected. However, if an Internet study is designed for the purpose of collecting new data from participants (e.g., survey questionnaires and Internet experiments), informed consent should be electronically obtained from participants. This type of informed consent form is usually represented as the first Web page that potential participants encounter in an online study. The use of the Internet for research purposes remains a matter of public debate, in part due to the National Security Agency spying scandal (Black, 2013) that blew up after the disclosure of massive surveillance of citizens' e-mail and phone conversations. More protections of privacy of such communications may ensue in the future.

Protecting Confidentiality of Data

One justification researchers use for keeping participants uninformed about their inclusion in a field study is that the data collected from such sources are essentially anonymous, with no possibility of personally identifying the persons who provided the responses. Of course, if video or other recording techniques are used that preserve identifying information about the participants, the data are not anonymous, and participants should be given the right to consent whether their data will be used. However, when data are recorded without any identifying information of any kind, any invasion of privacy is temporary, and confidentiality of the data is insured in the long run.

Even when research is not disguised, recording of participant identifying information to maintain confidentiality of data should be avoided whenever possible (Fisher & Vacanti-Shova, 2012). Assuring participants of the confidentiality of their responses is not simply for their benefit but can also increase the likelihood that they will be open and honest in their responses (Blanck et al., 1992; Boruch & Cecil, 1979). Esposito and colleagues (1984) compared responses on a personality test given by participants who had been assured their responses would be "strictly confidential" to those given by participants who had not been given such assurances. Those in the confidentiality condition provided data that were less influenced by social desirability biases than those in the control condition. This effect was obtained even though participants recorded their names on the tests they took.

Protecting confidentiality is relatively easy when no identifying information (such as names, social security numbers, etc.) is recorded with the collection of data from participants. In many cases, however, participants' names or other identifying features are recorded for various reasons, and under those circumstances protection of confidentiality can present legal or ethical dilemmas for researchers (Blanck et al., 1992). This can occur when the research involves sensitive information (e.g., testing for HIV) or potentially illegal or dangerous behavior (e.g., child

abuse) where reporting to partners or authorities may be seen as an ethical or legal responsibility. Research data in situations such as these are subject to subpoena, and researchers have sometimes been put into a painful conflict between their ethical responsibilities to research participants and their legal obligations. For research on some sensitive topics in the United States, it is possible to obtain a “certificate of confidentiality” from the Public Health Service (Sieber, 1992) that protects participant information from subpoena, but most research involving human participants is not protected in this way.

Datasets and archival research. Related to the general issues of invasion of privacy and confidentiality is the debate over creation of various national datasets for research purposes. Many public and national datasets available online, and those distributed by reputable organizations, have removed personally identifying information about the participants. Such datasets may be obtained online from the Pew Research Center, Census Bureau, Internal Revenue Service, Social Security Administration, and other agencies.

Longitudinal research might require that data records be kept about individual respondents, but these methods do not necessarily require that those individuals be personally identifiable. Systematic controls to screen researchers who wish to be granted access to data are designed to protect individual privacy and are not necessarily inconsistent with research aims. In their review of the potential benefits and risks associated with the creation of a national data center, Sawyer and Schecter (1968) proposed several standards of operation that would provide safeguards for privacy for participants who may be identified in a dataset. Among their major suggestions are the following:

- Only objective information be included in dataset
- Individuals be given the right to review their files for accuracy, and to have errors corrected
- Files of individuals be identified only by code numbers, with access to personal identification strictly restricted
- Security precautions be instituted for screening data users and restricting access

These last two suggestions concern the fact that some identification of participant records might be required for adding new information to existing records longitudinally, or for producing a file for review at the participant’s own request. To restrict access to the translation between a unique identifying number (e.g., a random number given to each participant) and personal identification (e.g., participant’s full name), some sort of “linking” system (involving a mediating step between access to a particular file and connection between file number and person’s name) should be employed. An example is that developed by the American Council on Education (Astin & Boruch, 1970). Creating linking systems of this kind can be expensive, but such costs balance the scientific usefulness of large data banks against the risks to individual privacy.

The Regulatory Context of Research Involving Human Participants

Our preceding discussion of ethical dilemmas is intended to convey the idea that there are no simple absolute rules for deciding whether a particular research practice or method is ethical. Rather, considerable judgment is involved in weighing the potential value of the research against potential stress or other costs to research participants. Except for obviously dangerous or damaging actions on the part of the researcher, ethical decision-making involves a cost-benefit analysis rather than the promulgation of absolute strictures and rules. Much of the responsibility for making this assessment falls on each scientist, but an individual researcher alone is not always the best judge of what is valuable and necessary research and what is potentially harmful to

participants. In fact, there is good evidence that biases enter scientists' assessments of the utility of their own research (Kimmel, 1991). For that reason, the conduct of research that meets reasonable ethical standards and procedures is not just a matter of personal judgment – it is the law.²

Almost all social and behavioral research in the United States that is supported by federal funds or conducted in educational or research institutions that receive federal funding (of any kind) is subject to federal regulations regarding the conduct of research involving human participants. The primary directive is 45CFR46 in the Code of Federal Regulations, known as the “Common Rule.” The code (last revised in 2018; for an up-to-date summary, see <http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html>) stipulates certain principles for protecting the welfare and dignity of human participants in research and prescribes policies and procedures that are required of institutions in which such research is conducted. As mentioned in Chapter 6 in connection with the initial planning stages of any research study, failure to comply with the procedures dictated by federal regulations can have dire consequences not only for the individual researcher involved but for the whole institution in which he or she works.

Institutional Review Boards

Much of the responsibility for complying with federal regulations is delegated to the institutions (e.g., universities) in which the research is conducted. Every institution in which federally funded research is carried out is required to set up an **institutional review board** (IRB), a committee that evaluates, approves, and monitors all research projects in an institution with respect to ethical requirements and practices. The IRB is appointed by the university administration but with certain requirements for representation by members of the community and legal experts, as well as scientists from departments across the institution. Before any program of research is begun, it is the principal investigator's responsibility to submit a complete description of the proposed research purposes and procedures to the IRB for review. Based on the information provided by the investigator, the members of the IRB evaluate the potential costs and risks to participants in the research as well as the potential benefits of the research if it is conducted as planned.³ A schematic representation of the IRB review process as it is practiced in most research universities is provided in Figure 21.1.

If the research described appears to meet ethical standards as set by the IRB committee, approval will be given to the investigator to conduct the research as described. Usually, approvals are granted for a maximum of 12 months; if the research has not been completed within that time, the project must be resubmitted for continuing approval. If the IRB does not feel that the researcher has provided sufficient information to assess the potential risks of conducting the study, or if the proposed procedures do not appear to be fully justified, the proposal will be sent back to the investigator with contingencies or changes that must be made before the research can be approved. In the majority of cases, the review process ultimately results in a project that can be approved for implementation. But on occasion, the IRB can, and will, refuse to allow certain research studies to be done because they are deemed to be unethical or excessively risky.

Although many scientists regard the IRB review process as cumbersome and time consuming, most recognize that it is now a standard part of the design and conduct of research involving human participants. Low-risk research that does not involve deception or issues of confidentiality usually can be handled by expedited review. But in other circumstances, most notably those involving potential danger, deception, or blatant manipulation of one form or another, the institutional review board serves a valuable function by requiring researchers to defend the legitimacy of the research, the necessity for the questionable practices, and the cost/benefit analysis involved in conducting the investigation.

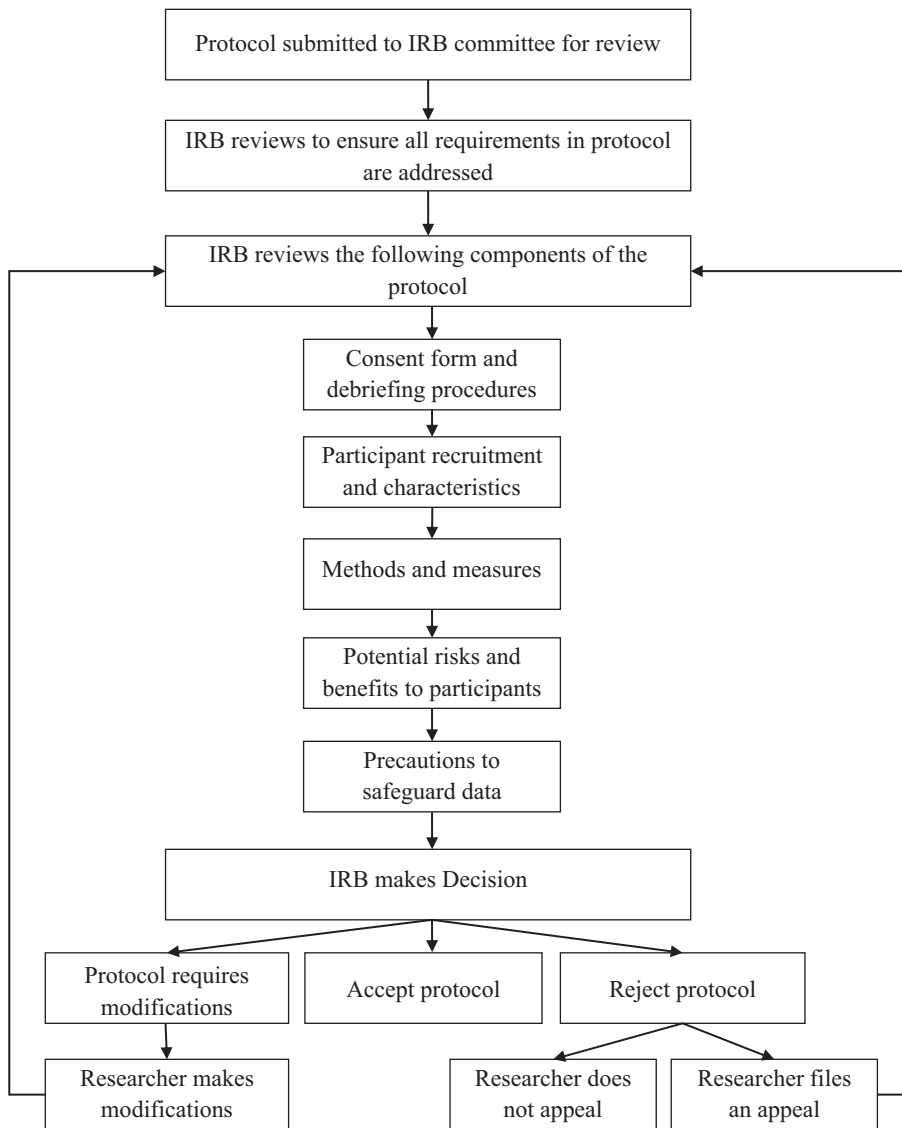


FIGURE 21.1 Process of IRB review.

An important feature of the institutional review board is that it typically does not consist solely of the researcher's colleagues – many of whom, perhaps, have planned or conducted research similar to that under consideration – but rather of a group of impartial individuals, scientists and laypersons alike, whose primary goal is the protection of participants' rights. As such, this body often can alert conscientious investigators to a potential problem that neither they nor their colleagues had noticed, simply because they were too involved with the technical and theoretical details of the research problem to notice the threats to participants' rights that the research might entail. The IRB review process plays an important role in assuring ethical responsibility in contemporary social research.

Codes of Ethics

Although clearance by the local IRB is a mandatory aspect of research that involves human participants, this review procedure does not absolve researchers from further responsibility for the ethical conduct of their research. In addition to the procedures dictated by federal regulations, behavioral researchers also are subject to codes of ethics promulgated by scientific societies such as the National Academy of Sciences (1995) and the American Psychological Association (1983, 1985, 1992, 2002; Knapp, 2011) in the United States. Such codes and principles provide guidelines for scientists in the planning, design, and execution of research studies. However, rules can always be circumvented, and even a formal code will be ineffective unless sanctioned by social support reflected in editorial policies of professional journals, funding decisions of foundations and federal agencies, and other sources of professional recognition. Thus, it is our view that the best guarantee of continued concern over ethical standards is the frequent airing of ethical issues in a way that ensures exposure to each new generation of researchers.

Methodology as Ethics

Some social scientists draw a sharp distinction between ethical and methodological issues in the design and conduct of behavioral research, but others think ethics and methodology are inextricably intertwined. Rosenthal and Rosnow (1984), for instance, promote the philosophy that sound research design is an *ethical imperative*, as well as a scientific ideal. Considering that participants' time, effort, and resources are involved in the conduct of any social research, they argue that researchers are ethically obligated to do only research that meets high standards of quality, to ensure that results are valid and that the research has not been a waste of participants' time. Rosenthal (1994) has gone so far as to suggest that IRBs should evaluate the methodological quality of research proposals as part of the cost/benefit analysis in their decisions about the ethicality of proposed research projects. Critics of this proposal (e.g., Parkinson, 1994; Sears, 1994) argue that ethical concerns and evaluation of the scientific merits of projects should be kept separate because they involve different expertise and different types of standards.

We are sympathetic to the general idea that the participants' investment of time and effort in research studies should be a factor in evaluating the costs of conducting research, and that there is some ethical obligation to be reasonably sure that the research project is worth doing before such costs should be incurred. However, we also agree that IRBs should not be in the business of evaluating the methodological quality of research proposals, beyond some very general evaluations of the justification for the project and the qualifications of the principal investigators. As we hope the previous chapters in this book have conveyed, the criteria for good methodology are neither static nor absolute. Multiple methodological approaches are needed to meet different criteria and purposes of research, so no one study is likely simultaneously to possess high internal validity, construct validity, and external validity. Rather than IRBs reviewing a study based strictly on methods to be undertaken, the value of a particular research project must be evaluated in terms of the contribution it will make to a body of research employing various methods, rather than as an isolated enterprise.

Ethics of Data Reporting

One arena in which ethical principles and scientific ideals converge is the prescription to report results of research and the methods by which they were obtained honestly and completely. Sadly, in all branches of science there are cases in which out-and-out fabrication of findings has been

uncovered. A few notorious instances of fraud committed in research have been discovered, across disciplines, ranging from plagiarism to falsification of data (Broad & Wade, 1982; Stroebe et al., 2012). Although sanctions varied, for a good number of these individuals, their researcher career came to a complete end after the discovery of scientific misconduct. Mechanisms that have been recommended to help identify fraudulent studies include whistleblowers who work in research laboratories, the peer-review process of journals, and analyses that assess whether reported results are so statistically improbable as to require a reanalysis of the original data by impartial experts.

Cases involving fraudulent data clearly violate both scientific and moral principles. However, there are other areas of reporting where the boundaries between ethical and unethical practice are not so clear-cut. Selective reporting of only some results of a study often occurs, and data from some participants are dropped from analyses if they are suspect in some way. Such practices can be justified to reduce unnecessary error in understanding and interpreting results of a scientific study, but these practices can be abused if used to distort the findings in the direction of reporting only what the researcher had hoped to demonstrate. To avoid such abuses, researchers need to use systematic criteria if dropping data from their analyses and to be scrupulous about reporting how these criteria were applied.

Researchers are expected to be honest about reporting results that do *not* support their hypotheses as well as results that do support their predictions. In addition, researchers need to be honest about what their hypotheses were in the first place. Sometimes, the results of a research study are unexpected. This should be valued as part of the research enterprise – if we only got expected results, there would be some question about whether there was any need to undertake the research in the first place! When unexpected findings are obtained, we usually can generate post hoc explanations about why results emerged that way. This is a valuable part of the research process – post hoc explanations become hypotheses for new research. However, in a research report, it is important to distinguish between interpretations of findings that are made after the fact and hypotheses that were made before the study began. If post hoc explanations are reported as if they had been hypothesized beforehand, this is a practice that Norbert Kerr (1998) labeled “HARKing” – *Hypothesizing After the Results Are Known*. In the long run, this practice could compromise the validity of published research by increasing the prevalence of studies containing results rife with Type I errors (Simmons et al., 2011).

Open Science and Transparency

In light of these concerns, scientific societies and journal editors in the social sciences have been reconsidering some of the data reporting practices in their disciplines and establishing new norms and guidelines to encourage more complete and open reporting. In 2014, the Association of Psychological Science, for example, announced new guidelines for submissions of research articles to its flagship journal *Psychological Science* (see http://www.psychologicalscience.org/index.php/publications/journals/psychological_science/ps-submissions). Among the new requirements is a checklist of research disclosure statements that are intended to enhance the reporting of methods. At the time manuscripts are submitted for consideration for publication in the journal, submitting authors must confirm the following for each study reported in the manuscript:

- all excluded observations and the reasons for exclusion have been reported in the Methods section
- all independent variables or manipulations have been reported

- all dependent variables or measures that were analyzed for the research study have been reported
- information on how the final sample size for each study was determined has been reported

These new requirements have been implemented to encourage open reporting and to ensure that full and honest disclosure of all aspects of a research study does not detract from the chances that a manuscript will be accepted for publication.

The move toward more transparency and openness in the conduct of scientific research is also reflected in the founding of the Center for Open Science which has issued ten policy standards (TOP guidelines) for promoting transparency and reproducibility of scientific research (Nosek et al., 2015). Scientific organizations and publishers are also encouraging researchers to *preregister* their planned research by recording their hypotheses (or research questions), methods, measures, and data analysis plans in advance of conducting the studies (Nosek et al., 2018; Soderberg, 2018).⁴ The commitment to such advance planning should not preclude analyzing and discussing unexpected results when the data are in, but it does reduce the possibility of HARKing by ensuring that we distinguish between prior expectations and ad hoc findings.

Ethical Issues Related to the Products of Scientific Research

The development and use of atomic power in the 1940s quite effectively exploded the myth that scientific research is immune from considerations of morality and social values. Since that time, scientists have had to come to grips with the issues of moral responsibility for the potential applications of their research discoveries. Some resolve this issue by rejecting all responsibility, claiming that scientific knowledge is essentially neutral, potentially serving good or evil depending upon decisions outside the scientist's control. Others feel that if scientists are in a reasonable position to foresee the immediate applications of their research efforts, then they must accept responsibility for the consequences of their continuing in that line of research.

This issue becomes most acute when the factor of research sponsorship is considered. When a research project is financed wholly or in part by some governmental or private agency, the researcher sometimes is obligated to report results directly – perhaps exclusively – to that agency. In such cases the purposes of the sponsoring agency will clearly determine at least the immediate application of information or technical developments derived from that research, and the scientist can hardly deny foreknowledge of such applications, whatever other potential uses the discovery may have. Given the growing costs of research in the physical and social sciences, more and more projects must rely on sources of funding other than those provided by such presumably impartial agencies as universities, and more and more scientists are facing a choice between abandoning a particular line of research or conducting it under the auspices of some private agency with special interests.

With respect to the long-range goal of social research – understanding human behavior in social settings – every researcher must be aware that as such knowledge accumulates, the potential for using it as a means of gaining control over other people also increases. Thus, the ethical considerations of any researcher in this area must include who will be privy to this knowledge in the long run, and what are the chances that it will come under the exclusive control of one segment of the social system (Kelman, 1968, 1972). Of more immediate concern is the current usage of information collected, or techniques developed, in the course of social research. For example, in research devoted to diagnosing antisocial attitudes or problematic personalities, various “disguised” or “projective” techniques have been developed, which purportedly assess the trait of interest under the guise of measuring something else. What is the responsibility of

the designers of these techniques when they are used by corporate personnel officers to weed out unsuspecting employees with potential anti-management values or attitudes? Or, alternatively, what is the responsibility of the researcher whose correlational study of social and attitudinal factors linked to student radicalism is used by university admissions officers to develop screening criteria for rejecting applicants?

The issue of social responsibility is made even more complex when one realizes that the conclusions to be drawn from research results or psychological tests often are grossly misperceived by naïve analysts. In the preceding example, for instance, the users of disguised tests or screening criteria might be largely unaware of the high degree of error (i.e., potential misclassification) evidenced in such selection devices when applied to individuals (see Fiedler & Grüning, 2021). Similar issues are raised with respect to research yielding controversial results that reveal differences (e.g., in intelligence or personality variables) between different ethnic or racial groups. Because the ethnic variable is inextricably confounded with a multiplicity of cultural and socioeconomic factors in contemporary Western society, the source of such differences in terms of genetic or cultural factors cannot be determined unambiguously, so researchers should report such results with extreme care with the necessary qualifications. Even so, there is no guarantee that other persons might not use the reported results to serve as justification for discriminatory practices based on the premise of innate differences between ethnic or racial groups.

Such potential misrepresentation of ambiguous research results has led some social scientists to suggest that a moratorium be declared on research involving race differences – either that such research not be conducted or that differences, if found, not be reported. Some scientists are horrified at the implication derived from this suggestion that research data should be withheld based on subjective moral judgments of individual researchers, whereas others take the more extreme position that because scientists may vary in what they consider ethically reprehensible or desirable, some commission should be formed to determine the distribution of research efforts and results. Neither of these suggestions appear consonant with the goals of science to learn and propagate valid information.

Conclusion

After reviewing these various ethical issues and associated policy suggestions, we have come to the position that the ethical dilemmas faced by the scientist-researcher cannot be solved entirely by any centralized decision-making body, which may place restrictions on the kind of research that can be undertaken or on the reporting of research outcomes. Rather, we feel that public interests will be served best by programs that actively promote alternative lines of research and competing theoretical (or philosophical) positions. To this end, we offer three suggestions:

- 1 That research programs that currently rely on exclusive sources of support instead be multiply sponsored, or receive support from a combined scientific research fund supported by budget allotments from several different agencies.
- 2 That to the maximum extent possible, all research reports, techniques, and summaries be made available for public distribution.
- 3 That emphasis be given to the social responsibility of individual scientists, or groups of scientists, to educate the public on the nature of their research results in ways that will enhance understanding of both the conclusions and the qualifications and limitations that must be placed on the generalization of those conclusion.

In this view, scientists would be encouraged to resist associating with research programs that prevent public access to scientific data and, similarly, to avoid placing the stamp of scientific

respectability on research that is seriously flawed, owing to methodological difficulties, or whose limitations are not clearly specified. These suggestions reflect our conviction that researchers who arm themselves with the mastery of diverse methods and apply this knowledge as active producers and consumers of open information are equipped with the best weapons against the misuse of scientific data or instrumentation.

Notes

- 1 They also, however, were somewhat more likely than nondeceived subjects to doubt the trustworthiness of psychologists!
- 2 Throughout this chapter we are describing regulations and procedures for ethics review that are based on federal laws and guidelines in the United States. The regulatory environment for the conduct of research with human participants varies from country to country, although many countries have adopted procedures that are modeled after those in the United States. Each researcher, of course, is responsible for knowing the legal requirements in any country where they (or their collaborators) are conducting studies.
- 3 Some research studies are considered “exempt” from review. These include secondary data analyses that use public databases or earlier (IRB-approved) studies have already been collected, and, as such, do not involve new use of human participants. Even under these conditions, many institutions still require IRB perusal, but usually these are perfunctory, and they are deemed exempt from further review.
- 4 One popular platform for preregistration and data sharing is Open Science Framework registries (<https://osf.io/registries>).

References

- American Psychological Association (1983). *Ethical principles in the conduct of research with human participants*. Washington, DC: APA Press.
- American Psychological Association (1985). *Standards for educational and psychological testing*. Washington, DC: APA Press.
- American Psychological Association (1992). Ethical principles of psychologists and code of conduct. *American Psychologists*, 47, 1597–1611.
- American Psychological Association (2002). *Ethical principles of psychologists and code of conduct*. Washington DC: APA.
- Asch, S. E. (1956). Studies of independence and conformity: A minority of one against a unanimous majority. *Psychological Monographs*, 70(9) (Whole No. 416).
- Astin, A. W., & Boruch, R. F. (1970). A “link” system for assuring confidentiality of research data in longitudinal studies. *American Educational Research Journal*, 7, 615–624.
- Baumrind, D. (1964). Some thoughts on ethics of research: After reading Milgram’s “behavioral study of obedience.” *American Psychologist*, 19, 421–423.
- Black, I. (2013). NSA spying scandal: what we have learned. *The Guardian*. Retrieved December, 2013. from <http://www.theguardian.com/world/2013/jun/10/nsa-spying-scandal-what-we-have-learned>
- Blanck, P., Bellack, A., Rosnow, R., Rotheram-Borus, M. J., & Schooler, N. (1992). Scientific rewards and conflicts of ethical choices in human subjects research. *American Psychologist*, 47, 959–965.
- Boruch, R. F., & Cecil, J. S. (1979). *Assuring the confidentiality of research data*. Philadelphia: University of Pennsylvania Press.
- Broad, W., & Wade, N. (1982). *Betrayers of the truth*. New York: Simon and Schuster.
- Brock, T. C. (1965). Communicator-recipient similarity and decision change. *Journal of Personality and Social Psychology*, 1, 6506–6554.
- Brock, T. C., & Becker, L. A. (1966). “Debriefing” and susceptibility to subsequent experimental manipulations. *Journal of Experimental Social Psychology*, 2, 314–323.
- Bryan, J. H., & Test, M. A. (1967). Models of helping: Naturalistic studies in aiding behavior. *Journal of Personality and Social Psychology*, 6, 400–407.
- Campbell, D. T. (1969). Prospective and control. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research* (pp. 357–382). New York: Academic Press.

432 Concluding Perspectives

- Charness, G., Samek, A., & van de Ven, J. (2022). What is considered deception in experimental economics? *Experimental Economics*, 25, 385–412.
- Esposito, J. L., Agard, E., & Rosnow, R. L. (1984). Can confidentiality of data pay off? *Personality and Individual Differences*, 5, 477–480.
- Feldman, R. E. (1968). Response to compatriot and foreigner who seek assistance. *Journal of Personality and Social Psychology*, 10, 202–214.
- Fiedler, K., & Grüning, D. J. (2021). A social psychological toolbox for clinical psychology. *Zeitschrift für Psychologie*, 229, 154–164.
- Fisher, C. B., & Vacanti-Shova, K. (2012). The responsible conduct of psychological research: An overview of ethical principles, APA Ethics Code standards, and federal regulations. In S. Knapp, M. Gottlieb, M. Handelsman, & L. Van de Creek (Eds.), *APA handbook of ethics in psychology*, Vol 2: *Practice, teaching, and research* (pp. 335–369). Washington, DC: American Psychological Association.
- Freedman, J. L., & Fraser, S. C. (1966). Compliance without pressure: The foot-in-the-door technique. *Journal of Personality and Social Psychology*, 4, 195–202.
- Holmes, D. S. (1967). Amount of experience in experiments as a determinant of performance in later experiments. *Journal of Personality and Social Psychology*, 7, 403–407.
- Holmes, D. S. (1976a). Debriefing after psychological experiments: I. Effectiveness of postdeception dehoaxing. *American Psychologist*, 31, 858–867.
- Holmes, D. S. (1976b). Debriefing after psychological experiments: II. Effectiveness of postdeception desensitizing. *American Psychologist*, 31, 868–875.
- Jung, A. F. (1959). Price variations among automobile dealers in Chicago, Illinois. *Journal of Business*, 32, 315–326.
- Kelman, H. C. (1967). Human use of human subjects: The problem of deception in social psychological experiments. *Psychological Bulletin*, 67, 1–11.
- Kelman, H. C. (1968). *A time to speak: On human values and social research*. San Francisco, CA: Jossey-Bass.
- Kelman, H. C. (1972). The rights of the subject in social research: An analysis in terms of relative power and legitimacy. *American Psychologist*, 27, 989–1016.
- Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2, 196–217.
- Kimmel, A. J. (1991). Predictable biases in the ethical decision making of American psychologists. *American Psychologist*, 46, 786–788.
- Knapp, S. J. (2011). *APA handbook of ethics in psychology*. Washington, DC: American Psychological Association.
- Latané, B. (1970). Field studies of altruistic compliance. *Representative Research in Social Psychology*, 1, 49–61.
- Lefkowitz, M., Blake, R. R., & Mouton, J. S. (1955). Status factors in pedestrian violation of traffic signals. *Journal of Abnormal and Social Psychology*, 51, 704–706.
- Middlemist, R., Knowles, E., & Matter, C. F. (1976). Personal space invasions in the lavatory. *Journal of Personality and Social Psychology*, 33, 541–546.
- Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, 67, 371–378.
- Milgram, S. (1964). Issues in the study of obedience: A reply to Baumrind. *American Psychologist*, 19, 848–852.
- Milgram, S. (1969). The lost-letter technique. *Psychology Today*, 3, 30–33, 66–68.
- Milgram, S. (1974). *Obedience to authority*. New York: Harper & Row.
- Miller, S. E. (1966). Psychology experiments without subjects' consent. *Science*, 152, 15.
- National Academy of Sciences (1995). *On being a scientist: Responsible conduct in research*. Washington, DC: National Academy Press.
- Nosek, B., et al. (2015). Promoting an open research culture. *Science*, 348(6242), 1422–1425.
- Nosek, B., Ebersole, C., DeHaven, A., & Mellor, D. (2018). The preregistration revolution. *PNAS*, 115, 2600–2606.
- Ortmann, A., & Hertwig, R. (1997). Is deception acceptable? *American Psychologist*, 52, 746–747.
- Page, S. (2000). Community research: The lost art of unobtrusive methods. *Journal of Applied Social Psychology*, 30, 2126–2136.
- Parkinson, S. (1994). Scientific or ethical quality? *Psychological Science*, 5, 137–138.

- Piliavin, I. M., Rodin, J., & Piliavin, J. A. (1969). Good Samaritanism: An underground phenomenon? *Journal of Personality and Social Psychology*, 13, 289–299.
- Resnick, J. H., & Schwartz, T. (1973). Ethical standards as an independent variable in psychological research. *American Psychologist*, 28, 134–139.
- Rokeach, M., & Mezei, L. (1966). Race and shared belief as factors in social choice. *Science*, 151, 167–172.
- Rosenthal, R. (1994). Science and ethics in conducting, analyzing, and reporting psychological research. *Psychological Science*, 5, 127–133.
- Rosenthal, R., & Rosnow, R. L. (1984). Applying Hamlet's question to the ethical conduct of research. *American Psychologist*, 39, 561–563.
- Ross, L., Lepper, M. R., & Hubbard, M. (1975). Perseverance in self-perception and social perception: Biased attributional processes in the debriefing paradigm. *Journal of Personality and Social Psychology*, 32, 880–892.
- Sawyer, J., & Schechter, H. (1968). Computers, privacy, and the national data center: The responsibility of social scientists. *American Psychologist*, 23, 810–818.
- Schaps, E. (1972). Cost, dependency, and helping. *Journal of Personality and Social Psychology*, 21, 74–78.
- Schwartz, R. D., & Skolnick, J. H. (1962). Two studies of legal stigma. *Social Problems*, 10, 133–142.
- Sears, D. O. (1994). On separating church and lab. *Psychological Science*, 5, 237–239.
- Sieber, J. E. (1992). *Planning ethically responsible research*. Newbury Park, CA: Sage.
- Silverman, I., Shulman, A. D., & Wiesensthal, D. L. (1970). Effects of deceiving and debriefing psychological subjects on performance in later experiments. *Journal of Personality and Social Psychology*, 14, 203–212.
- 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.
- Smith, S. S., & Richardson, D. (1983). Amelioration of deception and harm in psychological research: The important role of debriefing. *Journal of Personality and Social Psychology*, 44, 1075–1082.
- Soderberg, C. K. (2018). Using OSF to share data: A step-by-step guide. *Advances in Methods and Practices in Psychological Science*, 1, 115–120.
- Striker, L. J. (1967). The true deceiver. *Psychological Bulletin*, 68, 13–20.
- Striker, L. J., Messick, S., & Jackson, D. N. (1967). Suspicion of deception: Implications for conformity research. *Journal of Personality and Social Psychology*, 4, 379–389.
- Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science*, 7, 670–688.
- Surgeon General's directives on human experimentation (1967). *American Psychologist*, 22, 350–359.
- Walster, E., Berscheid, E., Abrahams, D., & Aronson, E. (1967). Effectiveness of debriefing following deception experiments. *Journal of Personality and Social Psychology*, 4, 371–380.
- Walther, J. B., Jang, J.-W., & Edwards, A. A. H. (2018). Evaluating health advice in a web 2.0 environment: The impact of multiple user generated factors on HIV advice perceptions. *Health Communication*, 33, 1, 57–67.
- Webb, E. J., Campbell, D. T., Schwartz, R. D., Sechrest, L., & Grove, J. (1981). *Nonreactive measures in the social sciences*. Boston, MA: Houghton-Mifflin.
- Weubben, P. L. (1965). Honesty of subjects and birth order. *Journal of Personality and Social Psychology*, 5, 350–352.

GLOSSARY

Actor A person who perceives, records, and provides information or a score regarding the interpersonal interaction with a partner in a dyad.

Actor-partner interdependence model A multivariate framework for examining two or more variables from each member of a dyad to determine the correspondence and reciprocity of features of their interpersonal relationship (love, trust, etc.).

Analogue experiment In these studies, participants respond directly to an artificial research context that has been designed to reproduce or mimic selected features of a real-world situation.

Apprehensive participants These participants are concerned that the experimenter will use their performance to evaluate their abilities, personality, social adjustment, etc., and react accordingly in the study.

Artifact An artificial or spurious finding, a “pseudo-effect” that results inevitably from the properties of the measuring instrument or from the method of data collection employed.

Ascendancy searching The reference list of a recent and relevant primary study is consulted to perform a backward search to find chronologically earlier relevant studies (ancestors).

Asymmetric dyad In these dyads, members are distinguishable by a relevant feature, such as their respective role or status in the interpersonal relationship.

Attention check A question inserted into experimental materials specifically to test whether participants are paying attention to instructions or questionnaire items.

Autocorrelated errors Measurement error that occurs if random events that affect the measure obtained at one time carry over and are correlated with measurements taken at temporally adjacent points in time.

Bidirectional causation A relation in which changes in A produce changes in B and, in addition, changes in B produce changes in A.

Bivariate correlation Applied to examine the association between two variables and includes the Pearson correlation, point-biserial correlation, and phi correlation.

Block design An experimental design that involves a factorial combination of experimental treatments and a nonexperimental variable.

Blocked randomization Participants are sorted into categories based on some individual difference variable and then randomly assigned to experimental treatments within blocks.

- Census** A survey or enumeration involving all the potential units of a target population (the universe of interest).
- Closed-ended question** An item on a questionnaire or a structured survey that specified the answers that may be chosen in response to a query (see also Open-ended question).
- Cluster sampling** In this form of sampling, geographic locations (or segments) are randomly sampled, and all members of the selected segment are used in the sample.
- Coding unit** The fundamental unit of interest to be attended to when performing a content analysis.
- Coding variable** The presence, frequency, or quantity of a coding unit to be recorded into a dataset and later used as a variable in statistical analyses.
- Coefficient of determination** The squared value of the Pearson correlation (r^2). It represents the proportion of variance shared between two variables.
- Coefficient of reproducibility** An indicator of the internal consistency of a Guttman scale.
- Cohen's d** The standardized difference between two means and should be used as an indicator of effect size if the relation involves a binary variable and a quantitative variable.
- Cohen's kappa** An index used to assess the extent of agreement between two coders of a qualitative categorical variable, which controls for chance agreement.
- Comparison time-series design** A design method that combines features of interrupted time-series and comparison group designs.
- Complete counterbalancing** Used to control for treatment order effects. Sequencing of treatments involves every combination of treatment orderings, and each participant is randomly assigned to one of the sequences.
- Conceptual hypothesis** A prediction about relations involving the theoretical constructs. It guides the purpose of a research study.
- Conceptual replication** An attempt to reproduce the results of a previous study by using different operational definitions to represent the same constructs.
- Concurrent validity** The extent to which a measure is related to, or explains, a relevant criterion behavior, with both variables assessed at the same occasion.
- Confederate** An actor or accomplice used by a researcher to pretend to be just another participant in a study.
- Confirmatory factor analysis** A statistical technique to assess the multidimensional structure of a scale. Used when the researcher has hypotheses regarding the number and types of underlying factors that will emerge in the analytic solution.
- Confound** A type of extraneous variable that affects the dependent variable in a way that cannot be separated from the effect of the independent variable, thereby posing a threat to internal validity.
- Constant** An attribute or feature that does not vary among the participants in a research study.
- Content analysis** A research method for classifying, evaluating, and studying recorded communications (e.g., audio, visual, images, multimedia) in a systematic and objective manner.
- Content validity** An estimate of the extent that a measure adequately represents the complete range or breadth of a construct under consideration.
- Context unit** In content analysis, this is the embedded context within which the meaning of a coding unit is to be inferred.
- Convenience sampling** In this weak form of sampling, the respondent sample consists of people most readily accessible or willing to be part of a study.
- Convergent validity** An estimate of the relation between measures of constructs that are theoretically related.

Criterion validity The extent that a measure is related to, or explains, a target outcome or criterion, usually a behavior.

Cronbach's alpha An index of the internal consistency (reliability) of a set of items in a scale.

Cross-lagged panel model Statistical method to determine the temporal direction of relations involving variables using a nonexperimental longitudinal data, and consisting of synchronous correlations, autoregression paths, and cross-lagged paths.

Debriefing A session following participation in a study, in which participants are informed of the true purpose of the research.

Deductive content analysis In this method, the researcher determines the coding categories and units that will be the basis of the coding system used in the analysis prior to conducting the coding.

Demand characteristics The totality of all social cues communicated in a laboratory not attributable to the manipulation, including those emanating from the experimenter and the laboratory setting, which influence or alter participants' responses.

Dependent variable A measured outcome or consequence, not manipulated by the researcher, and expected to be influenced by (or dependent upon) manipulation of the independent variable in an experiment.

Descendancy searching An important early and relevant primary study is identified and used as the starting point to search for all chronologically later studies (descendants) that have cited it.

Discriminant validity The extent that measures of constructs that are theoretically unrelated are independent of one another.

Disproportionate stratified (random) sample A stratified sample in which a different sampling fraction is used in each of the strata.

Double-barreled question A single item that asks more than one question at the same time.

Double-blind procedures A method in which both the participants and the researcher responsible for administering a study are made unaware of the experimental condition to which participants were assigned.

Dyadic variables Characteristics or features of interpersonal relationship between the two persons or their combined outcomes (e.g., level of warmth of their interaction, trust, etc.).

Ecological validity An aspect of external validity, referring to the extent to which an effect observed in a research study is typical or representative of a population.

Effect size A standardized metric of practical significance, it indicates the magnitude of relation between two variables, and serves as the main unit of analysis in a meta-analysis.

Endogenous variable A variable explained by determinant or predictor variable or variables, as postulated in a structural equation model.

Environmental manipulation These treatments entail the systematic manipulation of some aspect of the physical setting of a research study.

Epsem method A sampling design in which each eligible respondent has an equal probability of selection.

Evaluation research The application of social science methodology by program evaluators to assess effectiveness of social programs or interventions.

Event-related potential Electrodes attached to multiple regions of the scalp are used to measure these rapid waves of electrical brain activity that arise during the processing of information.

Exact replication An attempt to reproduce the results of a previous study by using the same procedures, particularly the same operationalizations, to represent the same constructs.

- Exogenous variable** A variable not explained by a determinant or predictor, as postulated in a structural equation model.
- Experience sampling method** In this technique, respondents carry a pager, text messaging device, computer tablet, or other electronic reminder device to respond to researcher-produced questions about where they are, what they are doing, thinking, or feeling at that moment.
- Experimental methods** Techniques of inferring causal relations, in which participants' actions are affected by the controlled manipulation of variables determined by the researcher.
- Experimental realism** The degree that an experiment has a real impact on participants during a study. Experimental arrangements of high realism induce participants to attend carefully to the task requirements.
- Exploratory factor analysis** A statistical technique to assess the multidimensional structure of a scale if the researcher does not have hypotheses regarding the number and types of underlying factors that will emerge in the scale's analysis.
- External validity** The extent of generalizability or certainty that results can be applied to other respondent groups, different settings, and different ways of operationalizing the conceptual variables.
- Face validity** A weak method of inferring validity based on superficial impressions regarding the extent that a measure appears to capture a construct.
- Factorial design** These designs involve combining levels of one independent variable with levels of all other independent variable(s) to expand the number of conditions in an experimental design.
- Faithful participants** Research participants willing to cooperate fully with almost any demand by the experimenter. They follow instructions scrupulously and ignore any suspicions they might have regarding the true purpose of the study.
- Field study** Research conducted outside the laboratory; participants are studied in their own naturalistic environment or context.
- File drawer problem** Unpublished studies often are tucked away in physical or electronic file drawers, and therefore ignored in a meta-analysis because they are not published. This results in a meta-analytic synthesis that is not representative of all the research that has been conducted.
- Fixed-effect model** In this meta-analysis approach, effect sizes of primary studies with larger sample sizes are weighted more heavily to calculate the summary effect.
- Fleiss' kappa** An index used to assess the extent of agreement between two or more coders of a qualitative categorical variable while controlling for chance.
- Focus group** An interview strategy in which a moderator obtains information by encouraging verbal discussion and communication about a focused topic among a group of participants.
- Forced choice measure** A closed item questionnaire or scale that requires participants to select an option listed.
- Functional magnetic resonance imaging (fMRI)** This is a measure of the relative degree of oxygen flow through blood vessels in the brain. It is used to identify which brain regions are implicated in various cognitive and emotional processing functions.
- Funnel plot** Primary studies are graphed in a scatter plot to inspect the distributional pattern of effect sizes and to evaluate the extent of publication bias in a meta-analysis.
- Good participants** A label applied to research participants who attempt to determine the experimenter's hypotheses and to confirm them.
- Group interview** In this method, an interviewer verbally requests information from more than one interviewee at the same time, but they are discouraged from interacting with each other.

- Group variables** The characteristics or variable features of a set of three or more persons (e.g., the age of the participants in a group, the total number of widgets created by a work team, etc.).
- Growth curve model** An analysis conducted to determine the longitudinal trajectory or shape of observations for participants measured at multiple time points.
- Hypothetico-deductive model** A method of generating hypotheses for scientific research by logical deduction from established theory.
- Implicit association test** An indirect measure that assesses automatic associations between mental concepts by requiring participants to make classification judgments as quickly and accurately as possible.
- Implicit responses** A person's responses to, or evaluations of stimuli, reflecting internal judgments of which they are not necessarily aware.
- Independent variable** A feature or characteristic manipulated independently of its natural sources of covariation to produce different conditions in an experimental study.
- Inductive content analysis** In this method, the researcher allows coding categories and units to emerge empirically by continually revising and updating the coding system during the coding process.
- Informed consent** An informational process by which potential participants are told what participation in a study will involve, so that they can make a reasoned decision about whether or not to engage in the research.
- Institutional review board (IRB)** A committee that evaluates, approves, and monitors all research projects in an institution with respect to ethical requirements and practices.
- Instructional manipulation** A treatment that involves modification of the description of the purposes and procedures that participants encounter in a study.
- Instrumentation error** A threat to internal validity, in which scores are affected by changes in the properties of the measurement instrument, rather than as an outcome of exposure to different levels of a treatment.
- Interaction effect** An outcome that is evident when the effect of an independent variable on a dependent variable is altered or moderated by variations in the level of other independent variable(s), while controlling for the main effects.
- Internal validity** Indicates the extent that inferences of causation can be made about the obtained relation between an independent variable and a dependent variable.
- Interrupted time-series design** A method in which the relative degree of change that occurs after a quasi-experimental treatment is examined by comparing observations across a series of time points prior to the treatment with observations taken at time points after its occurrence.
- Interval scale** A measure in which higher scores represent increasingly greater levels of a construct (e.g., weight, height, religiousness). In such measures, equal distances in numbers reflect equal gradations in the critical variable across all levels of the measurement scale.
- Interview** A data collection method in which participants verbally communicate information about their behavior, thoughts, or feelings in response to questions verbally posed by an interviewer.
- Intraclass correlation (ICC)** An index of interrater reliability for two or more coders of a quantitative variable used to assess the extent to which participants have more homogeneous scores within higher-order grouping units relative to variability of participant scores across all groupings in a multilevel model.

- Involuntary participants** Participants who feel that they have been coerced to spend their time in an experimental investigation and may vent their displeasure by actively attempting to ruin the study.
- Item response theory** A technique used to determine how each item in a scale operates in terms of difficulty, discrimination, and guessing to determine the overall score.
- Known groups method** A validation approach in which a measure is validated by its performance in linking scores with groups known (or strongly assumed) to be high or low on the measure.
- Latent factor** A factor that has removed measurement error using techniques such as structural equation modeling.
- Latent structural equation model** The result of the integration of confirmatory factor analysis and path analysis.
- Latent variable** A characteristic or construct that is not directly observable and therefore requires inferences to be made by the researcher.
- Leading question** Questions phrased in a way to suggest the expected answer; they involve a premise that must be accepted to answer the question (e.g., “When did you stop abusing your wife?” presupposes the individual at one time was abusive).
- Main effect** Occurs when overall mean differences are observed on a dependent measure as a function of one independent variable, while all levels of other independent variables are held constant.
- Manifest variable** A variable that is directly observable by the researcher.
- Manipulation check** A question or series of questions asked of participants to determine if they noticed the manipulation that was delivered in the experiment.
- Measurement construct validity** The overall validity of the measure with respect to the theoretical construct of interest.
- Measurement scale** The assignment of numbers to specific observations to reflect variations among those observations.
- Median split** A process in which participants are divided at the “middlemost” score of a variable, with 50% of participants falling above, and 50% falling below the split.
- Mediator** A third variable that serves as intermediary, to help explicate the chain of processes in a predictive relation.
- Meta-analysis** A quantitative approach used to summarize and facilitate interpreting primary studies that have addressed the same research question.
- Meta-moderator** A study-level characteristic in a meta-analysis to help determine whether a set of primary studies is significantly different from another set of studies in terms of the effect size.
- Missing at random (MAR)** This type of missing data occurs when the probability of missingness in a variable is not random, but may be fully explained by the other measured variables in the dataset.
- Missing completely at random (MCAR)** This type of missing data occurs if the probability of missing responses in a variable is not explained by (or is unrelated to) any other variable in the study.
- Moderator** A third variable that can either augment or dampen the cause–effect relation between variables.
- Monadic variables** These involve characteristics of single persons (e.g., age, religiosity, IQ).
- Mortality** A threat to internal validity that occurs when selection procedures, treatment differences, or other issues result in different proportions of participants dropping out of a study.

- Multidimensional scaling** A method that relies on ratings of the degree of similarity or dissimilarity among pairs of stimuli, to identify the number and types of dimensions that underlie the stimulus set.
- Multilevel model** An extension of multiple regression to estimate the relations of predictors on an outcome if the design involves a nested hierarchy of units.
- Multiple operationalization** An orientation that recognizes that no single operation or measurement provides sufficient information to adequately define a theoretical concept, so the construct is measured through several differing techniques or operations.
- Multiple regression** An extension of the Pearson correlation used to estimate the relations involving multiple predictors on an outcome.
- Multistage sampling** In this method, clusters of locations are sampled from a geographical sampling frame (as in cluster sampling), and then (unlike cluster sampling) units within each cluster are sampled as well.
- Multi-trait multimethod matrix** A validation technique in which multiple measures are used to assess the extent of association of theoretically related, but different, constructs, over and above the association that might come about simply because of shared measurement methods.
- Mundane realism** The degree to which various features of an experiment – instructions, treatments, and measurement operations – mirror real-world events that participants might encounter in their day-to-day experiences, rather than laboratory-specific operations.
- NAH** Shorthand for “not at home,” the bane of survey samplers.
- Narrative review** A qualitative approach for summarizing and interpreting primary studies that have addressed the same research question.
- Negative participants** Research respondents unhappy with their participation, who seek to understand the experimenter’s hypotheses so that they might sabotage the study.
- NHST** Null hypothesis significance testing; statistically assessing the probability that groups differ only by chance.
- Nominal group** A collection of individuals composed of the same number and types of people as a real group, but whose members work independently and do not interact. The group product thus is a combination of individual performances or outputs.
- Nominal scale** A measure requiring only that different numbers be assigned to observations, so they may be differentiated.
- Noncausal covariation** A relation in which changes in A are accompanied by changes in B, because both A and B are determined by changes in another variable, C.
- Nonexperimental methods** In this form of research, no variables are manipulated, but the relations of naturally occurring variables are measured.
- Non-random sampling** In this form of selection, a non-random mechanism is used to obtain a sample from a population.
- Nonvoluntary participants** Research respondents who do not feel their participation in a study was voluntary (see also Involuntary participants).
- Not missing at random (NMAR)** This form of missing data is obtained if a variable is not missing at random and its missingness is not explainable by the other measured variables in the dataset, but is explicable by other unmeasured variables.
- Null hypothesis** The possibility that there is statistically no difference among treatment conditions or no relation between measured variables.
- Observed score** The score obtained on a measurement instrument.
- Open-ended question** An item in a questionnaire with no constraint on allowable answers, rather than specifying specific response options and asking which one the respondent favors (see also Closed-ended question.)

- Operationalization** The process of translating an abstract theoretical construct or concept into a concrete specification of procedures and measures, so that it may be observed, recorded, and replicated.
- Ordinal scale** A measure in which the ordering of number labels corresponds to the rank ordering of observations along some dimension.
- Oversampling** In survey research, a process in which respondents who are underrepresented in the population are selected at greater rates than more commonly encountered respondents. This is done to ensure a stable representation of the oversampled group in the analysis.
- Pair comparison scaling** A stimulus scaling technique in which all possible pairs of stimuli are judged by respondents along some choice dimension (e.g., beauty, taste, sex appeal) stipulated by the researcher.
- Panel survey** A survey in which the same respondents are used repeatedly across survey administrations.
- Parallel forms reliability** A measure of internal consistency assessed by generating two different sets of items intended to assess the same underlying construct, which are then administered to the same participants at the same time and degree of relatedness calculated.
- Partial counterbalancing** A compromise to complete counterbalancing that uses a Latin square design to account for the fact that it may be cumbersome to assign participants to every possible sequence of treatment orderings.
- Participatory observation (participant observation)** A method that involves immersion of the researcher and social interaction between researchers and participants in the latter's milieu, during which time data, in the form of field notes, are unobtrusively and systematically collected.
- Partner** A person with whom an actor interacts during a dyadic exchange.
- Path model** A type of structural equation model in which predictive relations involving only measured variables are estimated.
- Pearson correlation** A bivariate correlation to examine the linear association between two quantitative variables.
- Phi correlation** A bivariate correlation to examine the association between two binary variables.
- Placebo effect** An outcome that occurs as a function of participants' beliefs in the efficacy of a treatment, rather than the active effect of treatment itself.
- Point-biserial correlation** A bivariate correlation to examine the association between a binary variable and a quantitative variable.
- Posttest-only control group design** An experimental design in which only a posttest, but no pretest, is given to randomly assigned participants in control and experimental groups.
- Power analysis** Estimating the sample size needed to detect a significant difference between groups at a specified effect size.
- Precision** A statistical value computed from members of a sample that serves as an inference, within some degree of sampling error, of a population value.
- Predictive validity** The extent that a measure is related to or explains a relevant criterion outcome that typically is assessed on a subsequent occasion.
- Pretest-posttest control group design** An experimental design in which both pretest and posttest are given to participants randomly assigned to control or experimental groups.
- Pretest sensitization** This threat to external validity occurs when participants in the experimental group are made sensitive to the treatment to which they are exposed as a function of prior exposure to the pretest.

Primary study An original study reporting the results of analyses of data collected from a sample of participants (e.g., a journal article, research report, dissertation).

Priming manipulation A treatment that induces a particular mindset or mental state of readiness in participants.

Propensity score This represents the conditional probability of an individual's membership in one group (e.g., the experimental group) over another (e.g., the control group), given the pattern of that person's responses on the covariates used in the score's development.

Propensity score matching A statistical technique that uses complex statistical procedures, statistically matching participants on as many covariates as can be specified by the researcher, to determine differences between comparison groups, or to create roughly comparable groups to be used in lieu of random assignment.

Proportionate stratified (random) sample A stratified sample in which the same sampling fraction is used in each of the strata.

Publication bias Refers to journal editorial practices to publish only statistically significant results, or findings in accord with the scientific zeitgeist of the time.

Quasi-experimental methods Variants of experimental methods that do not involve random assignment, but as in experiments, participants are exposed to some form of variable manipulation introduced by the researcher or social agents.

Question constraint A question format in which allowable answers are prespecified.

Question specificity Reflects the manner in which the question is phrased, acknowledging the fact that minor changes of wording may have marked effects on respondents' answers.

Questionnaire An instrument that typically involves a single item to assess each construct under study.

Quota sampling A method in which respondents are sampled non-randomly until a pre-defined number of participants for each of a specified set of subgroups (e.g., racial, ethnic, religious) are achieved.

Random assignment An allocation approach that requires that all persons available for a particular research study may participate in either the experimental or the control group, and that only chance determines the group to which any individual is assigned.

Random digit dialing A sampling approach that uses banks of phone numbers, randomly sampled from specified phone area codes, to construct a random sample of eligible respondents.

Random-effects model In this meta-analysis approach, effect sizes of primary studies are weighted as a compromise between sample size and number of primary studies used to calculate the summary effect.

Random error A factor affecting scores in a study that is attributed to chance events that tend to artificially widen the variability or spread of observed scores in a nonsystematic way.

Random sampling In this form of selection, a random mechanism is used to obtain a sample intended to be representative of the underlying population.

Random selection A form of creating a pool of potential research participants that requires that chance determine the selection of participants for a study, thereby helping to assure the generalizability of results from the sample to the entire population of relevant persons.

Rank order scaling This method involves respondents' ordering a set of stimuli along some dimension stipulated by the researcher.

Rating scale An instrument that makes use of multiple items to assess each construct, and typically is used by researchers with access to participants willing to take part in lengthy assessments.

- Ratio scale** The highest level of measurement, involving equal unit differences between adjacent numbers along all areas of a scale (as in interval scales), and a meaningful absolute zero-point.
- Recall measure** A participant's report of recollection of information previously presented.
- Recognition measure** A measure of a participant's memory of whether or not test material was presented previously.
- Regression discontinuity design** A quasi-experimental method conducted to test the existence of systematic relations between a pretest selection variable, used to place participants along some continuum (e.g., achievement, poverty) and a posttest measure of interest (e.g., school grades, job seeking).
- Regression equation** A mathematical expression used to denote the intercept and d coefficients of a regression model and is represented in the unstandardized format of $Y' = B_0 + B_1(X_1) + B_2(X_2) + B_3(X_3) \dots$ or the standardized format of $Z'_Y = \beta_1(z_{X_1}) + \beta_2(z_{X_2}) + \beta_3(z_{X_3}) \dots$
- Relations** How gradations of variations in one variable (or set of variables) predict or correspond to gradations or variations in another variable.
- Relevance** An issue of external validity, it has to do with the question of whether an effect obtained in a study is pertinent to events or phenomena that actually occur in the real world.
- Reliability** The consistency with which a measurement instrument assesses a given construct.
- Repeated-measures design** A statistical design in which the same participants are repeatedly measured on the dependent variable after receiving different treatment conditions.
- Research hypothesis** The empirical specification of a conceptual hypothesis regarding a testable directional prediction about specific relations in a study.
- Research question** A nondirectional question about specific relations in a study; less precise than a hypothesis, insofar as it merely queries the possibility of a relation, rather than specifying its direction or valence.
- Retrieval bias** Not all available primary studies that satisfied the inclusion criteria were included and synthesized in the meta-analysis. An exhaustive search by the meta-analyst is needed to avoid this problem.
- Robustness** An issue of external validity concerned with the extent an effect obtained in one research study can be replicated with different participants, operationalizations of the critical construct, and researchers.
- Role-playing simulation** A method in which participants are instructed to actively imagine that they are actors in a specified real-world situation and to respond as they believe they would in that context.
- Round robin design** A method used to examine multiple combinations of dyads in a group, by creating all possible pairs of persons, with each member rating each and every other member.
- Sampling error** The expected typical discrepancy between the estimates calculated from a sample and the value that would be obtained if the entire targeted population (i.e., a census) had been used.
- Sampling frame** A complete listing of the population from which members of a sample are drawn and used.
- Sampling weight** A means used to calibrate the demographic characteristics of a sample (e.g., age, race, sex, etc.) to that of the general population from which the sample was drawn.
- Scatter plot** A graph used to illustrate the correlation between two variables, with each point representing each participant's score on variable X and variable Y.

Scientific method A general approach for acquiring knowledge using systematic and objective methods to understand a phenomenon, which involves logic and data-checking feedback on the validity of results.

Selection In survey sampling, the process that guides the choice of respondents. Methods of selection play a major role in the extent to which survey results may be generalized to a larger population.

Selection-based interactions If participants were non-randomly selected to serve in the various comparison groups of a study, they may experience differences in history, maturation, testing, etc., which may produce differences in the final measurement, thus threatening internal validity.

Selection error A threat to internal validity that may occur if non-random processes are used to choose research participants and assign them to experimental and control groups.

Self-report A method in which respondents provide answers to a set of questions or scale items that inquire about their personal thoughts, feelings, or behaviors.

Semi-structured interview (structured-nonscheduled interview) An interview schedule in which a set of predetermined topics must be covered, but no specific questions are prescribed.

Sentence completion task An indirect measurement tool in which a partial sentence is presented to participants, who are required to complete the sentence.

Simple random sampling A sampling technique in which every member of the population in question has an equal (and nonzero) probability of being selected every time a unit is drawn for inclusion in a sample.

Snowball sampling A sampling technique in which initially selected respondents contact and recruit others in their social network for participation in the research.

Social manipulation An experimental treatment involving the delivery of scripted actions of another human being, usually a confederate working for the experimenter.

Social network analysis A method using dyadic relations as the basic unit of analysis, with the resulting data organized into matrix representations.

Sociometry The assessment of social acceptances and rejections of members in a given group; each member indicates liking or disliking each of the other members in the group.

Solomon four-group design An investigative approach combining posttest-only and pretest-posttest-control group true experimental designs to test the main effects of treatments, including pretest effects, and the interaction of treatments with the presence/absence of the pretest.

Spatial unit A measurement in which geographical, geometric, and dimensional properties (e.g., inches of newspaper column, dimensions of magazine article, shape of a photograph) are used as the units of analysis.

Split-half reliability A test of internal consistency reliability assessed by randomly dividing a scale into two sets containing an equal number of items, both administered to the same respondents simultaneously, with a test of relatedness calculated between these two summed scores.

Spurious An apparent relation between two variables that is actually an effect of some third variable.

Statistical power The probability of obtaining a statistically significant effect, if indeed that effect truly exists; dependent on the strength of the relation between variables and the number of observations (participants).

Statistical regression The tendency of extreme scores to regress to the mean of the distribution from which they were obtained. A threat to internal validity that arises as a function of measurement error coupled with the selection of participants on the basis of extreme scores.

- Statistical significance** A measure of the probability of obtaining an observed effect by chance, a threat to internal validity.
- Stimulus manipulation** An alteration of visual or verbal material in an experimental or quasi-experimental design.
- Stimulus scaling** Measurement technique designed to tap perceived differences among stimuli, rather than differences among participants.
- Stratified sampling** A technique in which a population of interest is divided into theoretically meaningful or empirically important strata. In pre-stratification, members are randomly drawn from each stratum (or subpopulation) before the sample is drawn. In post-stratification the classification process is done after data are collected from respondents.
- Structural equation model** A statistical technique that overcomes the limitations of a multiple regression analysis by allowing the researcher to estimate and test the strength of relations among multiple predictors and multiple criterion variables.
- Structured interview** A method in which a fixed list of questions is administered in the same order to all interviewees.
- Summary effect** The overall effect size across the primary studies of a meta-analysis; may be estimated using an unweighted, fixed-effect, or random-effects model.
- Symmetric dyad** In these two-person groups, members of each pair are indistinguishable by any theoretically relevant features such as their respective role or status in the interpersonal relationship.
- Systematic error (bias)** Consistent error in measures that tend to artificially inflate or deflate observed scores.
- Systematic sampling** A form of sampling requiring sampling every predetermined *n*th member from a population, after randomly choosing a starting point for the sampling process.
- Temporal unit** A measurement based on properties of time (e.g., minutes of a television or radio broadcast, seconds of a conversation).
- Test-retest reliability** A method of reliability testing assessed by administering a scale to the same participants twice, at different times. The degree of relatedness calculated between two administrations provides a measure of the test's stability.
- Thematic apperception test** An indirect measurement in which respondents view a deliberately ambiguous picture, and then generate a story about the characters and scenario depicted.
- Theory** A formulation based on observations and insights, consisting of a set of premises about why or how things occur or about the processes underlying those relations.
- Transcranial magnetic stimulation** A neuroscientific method to examine cause and effect between stimulations of a brain region and the consequent reaction.
- Triangulation** Converging on a phenomenon through multiple, diverse measures with non-overlapping sources of error.
- True experiment** A research study in which levels of the independent variable are manipulated by the experimenter and participants are randomly assigned to different experimental conditions.
- True score** The replicable feature of a concept that is being measured.
- Two-phase sampling** A sampling model in which a larger population (sometimes a census) is sampled, and then some respondents are selected to complete a second, more extensive survey.
- Type I error** An error of inference where a relation is detected due to chance, when it does not actually exist.

Type II error An error of inference where the null hypothesis is accepted when it is actually false and a true relation exists.

Unfolding technique A method to ascertain the dimensions underlying the preferences of a group of respondents to a set of judgment stimuli.

Unidirectional causation A relation in which changes in A produce subsequent changes in B, but changes in B do not influence A.

Unstructured interview (exploratory interview) A technique that encourages freely generated questions by interviewers and responses by interviewees; best used in the exploratory research phase.

Unweighted model A meta-analysis method in which effect sizes of primary studies are mathematically averaged to calculate the summary effect, without taking into consideration the sample size of each study.

Validity The degree of relation, or overlap, between a measurement instrument and the construct it is intended to assess.

Variable A characteristic or attribute that may differ, and is the basic unit of observation and measurement in a study.

Voluntary participants Respondents in a study who are aware that they are under investigation, but have made a conscious decision that the possible benefits involved outweigh the costs of being in the study.

Vote counting A literature review method that involves counting and comparing the number of primary studies that are statistically significant, or that support one explanation or another.

Z score A standardization procedure to ensure that all variables are on the same metric ($M = 0$, $SD = 1$), and is calculated by subtracting each person's score from the mean and dividing that by the pooled standard deviation.

AUTHOR INDEX

Note: **Bold** page numbers refer to tables; *italic* page numbers refer to figures and page numbers followed by “n” denote endnotes.

- Abbott, K. M. 385
Abelson, R. P. 38n3, 111, 117n3
Abrahams, D. 377, 420, 421
Acerbi, A. 145
Adair, J. G. 124, 127
Adam, E. 167
Adobor, H. 112
Adorno, T. W. 72
Agar, M. R. 141
Agard, E. 423
Aghakhani, N. 390
Agnew, N. M. 25
Ahern, G. 367
Ajzen, I. 207, 211
Alabastro, A. 143
Alexa, M. 283
Allison, S. T. 115
Altemeyer, B. 72
Altermatt, T. W. 387
Alvaro, E. M. 7, 85, 102, 163, 233, 343, 344
Alwin, D. F. 330n1
Ambady, N. 126
Ambrose, T. K. 156
American Psychological Association 310n3, 417,
427; Committee on Ethical Standards 417;
PsycINFO 402
American Psychological Association Task Force on
Socioeconomic Status 310n3
Amin-Esmaili, M. 293
Amso, D. 148
Amundson, D. R. 275
An, L. 322
Andersen, L. F. 144
Anderson, A. B. 294
Anderson, C. A. 7, 100, 138
Andrews, P. W. 298
Anthony, J. 137, 145
Antona, M. 112
Arabie, P. 342
Arcand, J.-L. 173
Arceneaux, K. 156
Argesti, A. 388
Argote, L. 388, 389
Aristotle 3
Armstrong, A. J. 54, 56
Arnold, S. 254
Aronoff, J. 387
Aronow, P. M. 156
Aronson, E. 25, 35, 36, 109, 132, 420, 421
Aronson, V. 377
Asch, S. E. 102, 103, 419
Aspinwall, L. G. 144
Astin, A. W. 424
Astrup, H. 144
Atalay, A. S. 351
Atkinson, P. 251
Auriacombe, M. 298
Austin, A. J. 358
Austin, J. 64
Austin, W. G. 270n1
Bacon, E. 322
Bakeman, R. 260, 264
Bales, R. F. 259, 261, 387
Ballard, J. 144
Balliet, D. 115
Baltes, M. M. 255
Banaji, M. R. 368

- Banks, W. C. 112
 Baraldi, A. N. 243
 Bargh, J. A. 104, 123, 149, 356, 357, 358, 361, 391n2
 Barnette, J. J. 329
 Barnlund, D. C. 386
 Baron, R. A. 6, 7
 Baron, R. M. 20, 21, 194, 195, 204
 Barros, M. M. 298
 Barry, C. L. 174
 Barsilou, L. 106
 Bartholomew, K. 292, 293
 Barton, A. H. 258
 Baruh, L. 121
 Basch, E. 243
 Battaglia, M. 143
 Battaglia, M. P. 240
 Baumberg, B. 239
 Baumrind, D. 419
 Bazarova, N. N. 389
 Beatty, P. 317
 Beck, L. F. 297
 Becker, H. S. 256
 Becker, L. A. 420
 Bedian, A. G. 49, 51, 52
 Beebe, T. J. 238
 Beecher, H. K. 126
 Bell, E. 174
 Bell, P. A. 6
 Bellack, A. 423
 Beltran, V. 298
 Bem, D. J. 111
 Bennett, B. L. 241
 Bennett, S. E. 297
 Ben-Porath, Y. S. 61
 Berelson, B. 285, 286
 Berenson, J. 139
 Berenson, R. 139
 Berent, M. K. 297, 317
 Bergan, D. E. 156
 Bergen, G. 297
 Berk, R. A. 172
 Berkman, D. 287
 Berkowitz, L. 101, 131
 Bernard, A. 112
 Berntson, G. 367
 Berntson, G. G. 364
 Berry, D. S. 205
 Berry, W. D. 171, 172
 Berscheid, E. 205, 420, 421
 Bessa, M. A. 298
 Best, J. 239
 Bettger, J. 385
 Beuckalaer, A. D. 150
 Bevan, J. L. 287
 Bhaskar, R. 4
 Bickel, L. 146
 Bickman, L. 26
 Biener, L. 343
 Biermann, M. 156
 Bird, A. 14
 Birnbaum, M. H. 146, 147, 149
 Bishop, G. R. 297
 Bizumic, B. 72
 Black, I. 423
 Blair, E. 307
 Blake, R. R. 422
 Blalock, H. M. 233
 Blanck, P. 423
 Blanton, H. 362
 Blascovich, J. 362, 364, 365
 Bliwise, D. L. 69
 Block, J. 72
 Bloom, P. N. 303
 Blumberg, S. J. 225
 Blume, L. E. 117n4
 Boden, M. T. 141
 Bodenhausen, G. V. 360
 Bodur, H. O. 351
 Boekeloo, B. O. 240
 Bogardus, E. S. 322
 Bogden, R. 251
 Bohus, M. 156
 Bolger, N. 255
 Bollen, K. A. 50
 Bommel, P. 112
 Bonaiuto, M. 155
 Bonnes, M. 155
 Bonnet, C. 298
 Boorman, S. 383
 Borden, R. J. 143
 Borenstein, M. 403, 405, 406, 415n4
 Borg, I. 335, 345n3
 Borgatta, E. F. 387
 Borgida, E. 258
 Borsboom, D. 48
 Boruch, R. 156
 Boruch, R. F. 423, 424
 Borzekowski, D. L. 288
 Boster, F. J. 137
 Both, S. 7
 Bourjolly, J. N. 308
 Bousfield, A. K. 370n5
 Bousfield, W. A. 370n5
 Bousquet, F. 112
 Box, G. E. P. 171
 Boyd, R. L. 322
 Boyle, K. 275
 Bradburn, N. M. 294, 307, 317, 330n1
 Bradford, J. 240
 Bradley, M. M. 365
 Bradley, M. T. 40
 Bradley, R. H. 174
 Brady, J. 144
 Brady, W. 398
 Brand, P. L. P. 241
 Bray, R. M. 381
 Brazier, G. 139
 Breau, D. L. 112
 Breen, R. L. 308
 Brehm, J. W. 111

- Breiger, R. 383
 Brennan, A. 239
 Brenner, L. A. 238
 Breuning, M. 233
 Brewer, M. B. 19, 25, 35, 103, 105, 115, 116, 130, 132, 140, 251, 353, 380
 Bridges 286
 Bridgland, V. M. E. 244n8
 Briesch, A. M. 48
 Broad, W. 428
 Broadfoot, A. A. 48
 Brock, T. C. 420, 422
 Brook, B. 112
 Brooks, J. O., III 401
 Brown, S. C. 370n5
 Brown, T. A. 50, 52, 69
 Brubaker, P. 275
 Brunner, G. A. 303
 Brunswik, E. 131
 Bryan, J. H. 422
 Bryant, F. B. 62
 Bryant, J. 36
 Buchanan, T. 146
 Buckhart, B. R. 61
 Bugental, D. B. 144
 Buhi, E. R. 242
 Buhrmester, M. 146, 148
 Bullock, J. G. 317
 Burger, G. 7
 Burgoon, M. 46, 85
 Burke, T. 287
 Burnette, C. B. 241
 Burrows, L. 123, 358
 Burtea, T. 69
 Busam, B. 145
 Bushman, B. J. 121, 138
 Byrne, D. 204, 205; *The Attraction Paradigm* 204

 Cabooter, E. 329
 Cacioppo, J. T. 130, 131, 146, 354, 364, 365, 367
 Cacioppo, S. 146
 Cain, P. S. 295
 Calderwood, L. 241
 Caldwell, B. M. 261, 263
 Callinan, S. 317
 Calne, D. B. 126
 Cameron, L. 116
 Campbell, D. T. 10, 19, 22, 27, 35, 61, 66, 67, 68, 69, 72, 73, 79, 120, 143, 144, 156, 160, 161, 163, 173, 322, 349, 380, 386, 402, 422
 Campbell, J. P. 8
 Campbell, L. 398
 Campbell, W. 385
 Canary, H. E. 145
 Cancelliere, C. 241
 Cannell, C. F. 296, 298, 301, 303, 307, 310n4
 Cantor, N. 355
 Carey, B. 397
 Carlopio, J. 124
 Carlsmith, J. M. 109, 111
 Carlson, J. 254
 Carlson, R. 111
 Carroll, J. B. 342
 Carroll, S. J., Jr. 303
 Carson, R. T. 297, 317
 Carter, M. 389
 Carter, T. J. 105
 Cartwright, A. 303
 Casey, M. A. 308
 Casey, R. L. 264
 Casler, K. 146
 Casline, E. 241
 Cassidy, C. 139
 Cassisi, J. E. 322
 Catanzaro, S. J. 69
 Cattaneo, Z. 367
 Cattell, R. B. 46
 Caul, W. F. 380
 Cecil, J. S. 423
 Cedraschi, C. 241
 Celentano, D. 156
 Cemalcilar, Z. 121
 Chabris, C. F. 148
 Chafouleas, S. M. 48
 Chalem, E. 298
 Chamberlin, P. E. 287
 Chan, D. M-L. 293
 Chandler, J. 241
 Chang, L. 68
 Chapman, L. J. 72
 Chapple, C. 69
 Chapple, E. D. 307
 Charles, E. P. 49
 Charness, G. 419
 Chartrand, T. L. 104, 356, 358, 361
 Chatterjee, G. 148
 Chaudhuri, A. 116
 Chave, E. L. 318, **320**
 Chen, M. 123, 358
 Chen, Q. 204
 Chen, X. 344
 Cheng, K. 139
 Cheng, T. L. 240
 Cherkasova, E. 141
 Cheung, C. 145, 146
 Cheung, G. W. 72
 Chico, E. 73
 Choi, I. 409
 Choi, W. 145, 146
 Chrisman, B. 145
 Christodoulou, J. 412
 Chu, V. 146
 Chung, S. 234
 Church, A. H. 238
 Cialdini, R. B. 7, 143
 Cicourel, A. V. 255
 Cillissen, A. H. N. 385
 Civic, D. 240
 Clark, J. C. 112
 Clark, R. D. 397

- Clarke, C. 361
 Clayton, R. B. 145
 Cliff, N. 322
 Clifton, A. D. 385
 Clore, G. L. 70, 381
 Coaley, K. 54, 55
 Cochran, W. G. 238
 Cohen, A. R. 111
 Cohen, B. D. 287
 Cohen, J. 24, 304
 Cohen, J. A. 98, 116n1, 265, 412
 Cohen, S. P. 261
 Coles, M. 367
 Collins, M. A. 144
 Collip, D. 345
 Combourieu, I. 298
 Compitello, S. 251
 Conaway, M. 297, 317
 Condon, J. W. 204, 205
 Condon, L. 73
 Connelly, B. S. 68
 Conolley, E. S. 111
 Conrad, F. 317
 Conrey, F. 361
 Contractor, A. A. 241
 Converse, P. E. 317
 Cook, D. A. 238
 Cook, S. W. 254
 Cook, T. D. 35, 61, 120, 124, 156, 163, 173
 Cook, W. L. 374, 375, 378, 379, 386
 Coolidge, F. L. 279, 287
 Coombs, C. H. 318, 335, 341
 Cooper, H. 403, 413, 415n4
 Cooper, H. M. 399, 414, 402
 Cooper, J. 7, 104
 Cooper, R. E. 335, 340
 Correll, D. N. 101
 Corso, M. 241
 Cortez, V. L. 144
 Cote, P. 241
 Cotter, K. N. 255
 Cotter, P. R. 304
 Coulter, P. B. 304
 Couper, M. P. 233, 317
 Courtney, D. 143
 Cox, J. 112
 Cox, W. T. L. 358
 Craik, F. 367
 Crandall, C. S. 398
 Crano, W. D. 7, 8, 10, 46, 55, 72, 102, 123, 130, 143, 147, 155, 156, 163, 171, 204, 205, 217, 233, 250, 281, 318, 335, 340, 343, 344, 391n2, 408
 Crites, S. 367
 Cronbach, L. J. 31, 45, 48, 49, 50, 57n2, 120, 202, 314, 323, 329, 391n1
 Cronkite, R. C. 111
 Crossley, H. M. 295
 Crowther, B. 387
 Csikszentmihalyi, M. 255
 Cuccolo, K. 397
 Cuddy, A. J. C. 343
 Cunningham, S. J. 360
 Cunningham, W. A. 368
 Currow, D. C. 141
 Curry, T. J. 380
 Cusack, R. 148
 Cuthbert, B. N. 365
 Cycyota, C. S. 238
 Daft, R. L. 8
 Dalal, R. 138
 Daneshfar, A. 112
 Daniels, W. L. 238
 Danielson, W. A. 279, 280
 Darley, J. M. 100, 113
 Davidson, R. 308, 367
 Davies, M. N. O. 52
 Dawes, R. M. 333, 341
 Dawson, E. J. 46
 Day, D. V. 49, 51, 52
 de Ayala, R. J. 54
 de Champlain, A. F. 52
 De Dominicis, S. 155
 de la Fuente-Fernandez, R. 126
 De Neve, K. M. 7
 De Simone, J. A. 147
 DeAndrea, D. 137, 145
 Deandrea, D. C. 389
 DeAngelo, L. 54, 57
 DeHaven, A. 429
 Delle, F. A. 145
 DeLongis, A. 254
 Deluse, S. R. 173
 DeMars, C. 54
 Dembitzer, L. 69
 Denis, C. 298
 Dennis, A. R. 390
 DePaulo, B. M. 260, 263, 265
 DeShon, R. P. 48
 DeSimone, A. J. 147
 Desjarlais, M. 377
 Deters, F. G. 145
 Deutsch, M. 343
 DeVellis, R. B. 52
 Devine, P. G. 358
 DeWall, C. N. 7
 DeWalt, B. R. 251
 DeWalt, K. M. 251
 Diehl, M. 390
 Dienstbier, R. A. 364
 Dietze, P. 317
 Dijkstra, W. 294
 Dillman, D. A. 293
 Ding, C. S. 345
 Dion, K. K. 205
 Dohrenwend, B. S. 302, 305
 Dollard, J. 6
 Donaldson, C. D. 105, 147, 213
 Donaldson, W. 356
 Dong, Y. 345

- Donnerstein, E. 131
 Donoghue, C. 322
 Doob, L. W. 6
 Dovidio, J. 116
 Dovidio, J. F. 349
 Downing, J. 317
 Drake, K. M. 167
 Drier, E. M. 241
 Duchaine, B. C. 148
 Duckitt, J. 72
 Dueck, A. C. 243
 Dull, V. 353
 Dulles, John Foster 287
 Dunlap, K. 145
 Dunn, E. W. 368
 Dunton, B. C. 360, 369
 Durlauf, S. N. 117n4
 Dutt, P. K. 339
 Duval, S. 412
 Dykema-Engblade, A. 388, 389
 Dyslin, C. W. 82

 Ebersole, C. 429
 Ebesutani, C. K. 329
 Edens, J. F. 112
 Edlund, J. 397
 Edwards, A. A. H. 423
 Egger, M. 411
 Eid, M. 68
 Einstein, Albert 397
 Ekman, P. 365
 Elliot, A. J. 255
 Ellis, J. M. 240
 Ellis, P. D. 413
 Ellison, J. W. 286
 Embretson, S. E. 48
 Emerson, R. M. 380
 Enders, C. K. 243
 Engle, K. M. 101
 Epstein, J. A. 260, 263, 265
 Epstein, J. S. 127
 Erber, R. 388
 Erkal, N. 116
 Erlebacher, A. 161
 Ernst, J. M. 364
 Escue, M. 342
 Esposito, J. L. 423
 Estes, W. K. 38n3
 Etcoff, N. 355
 Evans, D. C. 147
 Eveland, W. R. 385

 Fabrigar, L. R. 294, 315
 Fair, J. 156
 Fallon, M. T. 141
 Fan, X. 48, 57
 Fang, X. 298
 Fatseas, M. 298
 Faust, K. 384
 Fazio, R. H. 349, 360, 369

 Feeney-Kettler, K. A. 69
 Feldman, R. E. 422
 Fennis, B. M. 307
 Ferguson, M. J. 105
 Ferguson, R. W. 335, 345n6
 Ferrando, P. J. 73
 Festinger, L. 7, 111, 252, 253
 Fiedler, K. 398, 430
 Fielding-Smith, S. F. 255
 Fielstein, E. M. 141
 Fishbein, M. 207, 211
 Fisher, C. B. 423
 Fiske, A. P. 409
 Fiske, D. W. 66, 67, 68, 69
 Fiske, S. T. 343, 350, 353, 355
 Fleck, R. A. 112
 Fleiss, J. L. 266, 267
 Flora, C. 32
 Fode, K. L. 126, 127
 Forehand, R. 254
 Forgas, J. P. 139
 Forkus, S. R. 241
 Forscher, P. S. 358
 Forsman, L. 72
 Forsyth, D. R. 387
 Fortner, C. K. 174
 Fournier, E. 240
 Fowler, F. J. 310n4
 Fowler, L. R. 279
 Fowler, S. M. 112
 Fox, B. 342
 Francis, K. 258
 Frankel, M. R. 240
 Fraser, S. C. 141, 303, 422
 Freedman, J. L. 111, 141, 303, 422
 Freeman, S. 143
 Frenkel-Brunswik, E. 72
 Freud, Sigmund 8
 Friberg, O. 329
 Friesen, W. V. 365
 Frost, R. O. 69
 Fuchs, M. 239
 Funayama, E. S. 368

 Gable, S. L. 255
 Gaedde, L. A. 238
 Gaertner, S. L. 116
 Galanter, E. 342
 Galea, S. 174
 Galtier, F. 112
 Galton, F. 156, 157
 Gamson, W. A. 112
 Gangadharan, L. 116
 Gangestad, S. W. 298
 Ganucci Cancellieri, U. 155
 Garcia, D. J. 147
 Garcia, D. M. 147
 Gardiner, E. 361
 Gardner, C. 401
 Gardner, W. 105, 367

452 Author Index

- Gardner, W. L. 103
Garg, S. 391n3
Gatenby, J. C. 368
Gawronski, B. 361, 362
Geeza, A. A. 68
Geller, E. 265
Geller, S. 240
Genovese, Kitty 113, 114
Gerard, H. B. 36, 111
Gergen, K. J. 120
Germine, L. 148
Geuens, M. 73
Gilbert, D. T. 361, 409
Gillmore, M. R. 240
Ginos, B. 243
Ginsberg, E. 351
Girandola, F. 303
Glaser, B. G. 257
Glaser, B. G.: *The Discovery of Grounded Theory* 257
Glass, D. 144
Glass, G. V. 401, 402
Glasser, G. J. 239
Glathe, H. 356
Gleser, G. C. 48
Glick, P. 343
Goethals, G. R. 204
Goldinger, S. D. 335, 345n6
Goldman, M. S. 298
Goldman, R. M. 281
Goldstein, S. C. 241
Goncharenko, S. 241
Gonsalkorale, K. 145
Gonzalez, A. V. 7
Goodbey, G. 254
Goodie, A. S. 385
Goodman, A. 167
Goodson, P. 242
Gore, J. C. 368
Gosling, S. 144, 145
Gosling, S. D. 146, 148, 149
Gottman, J. M. 264, 388
Gounder, M. M. 243
Graef, R. 255
Graham, J. 146
Graham, J. W. 242
Graham, L. 144, 145
Grandpre, J. 85
Green, B. F., Jr. 322
Green, D. P. 156
Greene, D. 4
Greene, M. F. 167
Greene, P. 156
Greenfield, T. K. 240
Greenwald, A. G. 15, 361
Greenwald, S. 401
Greenwood, J. D. 110, 111
Greenwood, K. E. 255
Greifeneder, R. 145
Griffin, J. M. 238
Grill, J. P. 238
Grimm, K. J. 243
Groom, C. 361
Grove, J. 143, 144, 422
Groves, R. M. 238, 239, 293, 303
Gruning, D. J. 430
Gschwendner, T. 362
Guba, E. G. 256
Guglielmi, R. S. 362, 367
Guilford, J. P. 337, 340
Guinsburg, R. 298
Gulliksen, H. 40
Gupta, R. D. 40
Gutek, B. A. 293
Guttman, L. 320–322, 323, 325, 327
Gynther, M. D. 61
Haber, N. 145
Hacker, J. 390
Hackett, E. 146
Hadland, S. E. 235
Hahn, S. 251
Hahs-Vaughn, D. L. 239
Hajebi, A. 293
Halberstadt, J. B. 345
Hales, A. H. 145
Hallahan, M. 23
Hallberg, K. 173
Hallgren, K. A. 267
Hallmark, B. W. 387
Hambright, B. L. 287
Hamilton, Alexander 278
Hamilton, D. L. 104, 350, 355, 370n5
Hammersley, M. 252
Hampton, K. 385
Hancock, G. R. 48
Handali, J. 390
Handren, L. 55
Hanemann, W. M. 297, 317
Haney, C. 112
Hannula-Bral, K. A. 391n2
Harms, P. D. 147
Harris, A. M. 238
Harris, M. J. 126, 127
Harrison, D. A. 238
Hartway, J. 240
Harvey, A. M. 279, 287
Harvey, J. 289
Haselton, M. G. 298
Hassin, R. R. 105
Hastie, R. 117n3
Hatch, E. E. 167
Hatchett, S. 304
Hauser, D. J. 146
Hawkey, L. C. 364
Hawley, S. 322
Hayes, A. F. 20, 205
Hayward, M. 255
Hazan, C. 234
Heck, R. H. 199
Hedges, L. V. 403, 405, 406, 413, 415n4

- Hefazi, M. 293
 Heider, F. 382
 Heine, S. J. 121
 Hemovich, V. 233
 Hemphill, K. J. 254
 Henard, D. J. 54
 Henderson, A. J. Z. 292, 293
 Henle, M. 254
 Henrich, J. 121
 Henry, G. T. 173, 174, 226, 233
 Heon, N. 243
 Herbst, A. L. 167, 168
 Herman, J. B. 293
 Herrmann, D. 317
 Hertwig, R. 418
 Heyns, R. W. 264
 Higbie, C. E. 287
 Higgins, E. T. 356, 357
 Higgins, J. P. 403, 405, 406, 415n4
 Higginson, I. J. 141
 Hill-McManus, D. 239
 Hinsz, V. B. 388
 Hixon, J. G. 361
 Hoare, K. J. 258
 Hodge, R. W. 295
 Hoff, R. A. 141
 Hoffman, D. M. 172
 Hoffmann, W. 362
 Hoffmire, C. A. 238
 Hofling, V. 68
 Hofstede, G. J. 112
 Hogg, M. A. 143, 388, 389
 Hohman, Z. 155
 Hol, A. M. 56
 Holbrook, A. 297
 Holbrook, A. L. 317
 Holland, P. W. 382
 Hollingshead, A. B. 389
 Hollingshead, A. de B. 295
 Hollister, R. G. 156
 Holmes, D. S. 420
 Holmes, J. 239
 Holsti, O. R. 285, 287
 Hong, Y. 298
 Hoover, R. N. 167
 Hoppe, M. J. 240
 Hornik, R. 298
 Horowitz, I. L. 111
 Horwich, P. 14
 House, P. 4
 Hout, M. C. 335, 345n6
 Hovanitz, C. 61
 Hovland, C. I. 160
 Howard, W. D. 250
 Huang, C. 141
 Hubbard, M. 420, 421
 Hubbell, M. B. 254
 Hugenberg, K. 361
 Hulbert, H. 238
 Hulin, C. L. 8
 Hum, N. J. 287
 Hunt, E. B. 278, 285
 Hunter, J. E. 38n3, 49
 Hunter, S. 362
 Huo, D. 167
 Hussain, J. A. 141
 Hwang, H.- S. 261
 Hyer, M. 167
 Iida, M. 255
 Ipeirotis, P. G. 241
 Irgens, M. 397
 Isen, A. M. 141
 Itkin, S. M. 381
 Ito, T. A. 364, 365
 Itzhakov, G. 255
 Iyer, R. 146
 Jaccard, J. 6, 362
 Jackson, D. N. 72
 Jackson, J. R. 360, 369
 Jacobs, J. 269
 Jacobs, R. C. 386
 Jacobson, L. 126
 Jacoby, J. 6
 Jaffe, D. A. 112
 Jamieson, J. P. 255
 Jang, J.- W. 423
 Janis, I. L. 111, 144
 Jastro, J. 126
 Jemmott, J. B., III 156
 Jemmott, L. S. 156
 Jenkins, G. M. 171
 Jensen-Doss, A. 241
 Jilke, S. 156
 John, O. P. 121, 148, 149, 351
 Johnson, C. 390
 Johnson, D. R. 240
 Johnson, J. E. 144
 Johnson, K. M. 146
 Johnson, M. J. 141
 Johnson, P. B. 298
 Jones, C. R. 356, 357
 Jones, J. S. 251, 252
 Jones, S. 147
 Jones, S. P. 7
 Jordan Jackson, F. F. 145
 Jorgensen, T. D. 48
 Joyce, James: *Ulysses* 286
 Judd, C. M. 386
 Jung, A. F. 422
 Junne, F. 244n8
 Kaasa, S. 141
 Kahn, J. H. 199
 Kahn, R. L. 293, 303
 Kahneman, D. 4
 Kalantarian, H. 145
 Kalton, G. 228, 231, 238, 243n3, 295, 297, 300
 Kameda, T. 388

454 Author Index

- Kanazawa, Y. 317
Kapur, S. 367
Kardes, F. R. 360
Karl, K. A. 390
Karp, D. 251
Kashy, D. A. 260, 263, 265, 375, 378, 380, 386
Kasin, B. M. 144
Kasprow, W. J. 141
Katon, J. 238
Katz, I. R. 141
Katz, L. B. 104, 355, 370n5
Keener, J. 241
Keeter, S. 239
Kelloway, E. K. 49, 51, 52
Kelman, H. C. 111, 418, 420, 429
Kelsey, R. M. 364
Kendrick, R. V. 307
Kennedy, C. 239
Kennedy, J. L. 34
Kenny, D. A. 4, 20, 21, 69, 194, 195, 204, 211, 322, 374, 375, 378, 379, 380, 381, 386, 387, 391n3
Kerr, N. L. 387, 428
Kerr, P. 241
Kerr, T. 235
Kershaw, D. 156
Kerwin, J. 317
Kessler, R. C. 234
Kettler, R. J. 69
Kezer, M. 121
Kidder, L. H. 270n5, 349
Kiers, H. A. L. 69
Kiesler, S. 144, 149, 389
Killackey, E. 345
Kim, H. 365
Kim, J. 137, 145
Kim, S. 261
Kimmel, A. J. 425
Kinsey, A. C. 297, 299, 305
Kirkendol, S. E. 260, 263, 265
Kirkham, N. 148
Kish, L. A. 225
Kitayama, S. 409
Kjaervik, S. L. 121
Klein, D. 302, 305
Klein, R. 398
Klein, R. D. 112
Kleindienst, N. 156
Kleinman, S. B. 385
Klimoski, R. 389
Klinc, M. 48
Kline, A. 145
Kluegel, J. R. 293
Klumb, P. L. 255
Knapp, S. J. 417, 427
Knoke, D. 384
Knopov, A. 174
Knowles, E. 422
Kochovska, S. 141
Kohler, H. 385
Kolakowsky-Hayner, S. A. 54, 56
Kommer, D. 71
Kondracki, N. L. 275
Koo, J. 126
Kopp, R. J. 297, 317
Koppelman, N. F. 308
Kowai-Bell, N. 362
Kraemer, H. C. 401
Krambeck, H.-J. 115
Kramer, R. M. 115
Krantz, D. S. 144
Krantz, J. H. 138, 144
Krauskopf, P. B. 308
Kraut, R. 149
Krauth-Gruber, S. 106
Kreindler, D. 143
Krippendorff, K. 284, 285
Krishnan, R. 389
Krishnamurty, P. 317
Kroesbergen, H. T. 298
Kroner, D. G. 298
Krosnick, J. A. 233, 294, 297, 315, 316, 317, 328, 330n1
Krueger, A. B. 156
Krueger, R. A. 308
Kruger, J. 244n8
Krull, D. S. 361
Kruskall, J. B. 342
Kubey, R. 255
Kuder, G. F. 57n2
Kuhn, T. S. 4, 14, 397
Kuklinski, J. H. 384
Kulik, J. 139
Kundur, D. 143
Kuntsche, E. 317
Kunz, T. 239
Kwang, T. 146, 148

La Voie, L. 380, 381
Labouvie, E. 156
LaBrie, R. 240
Lac, A. 8, 55, 101, 105, 143, 211, 213, 233, 274, 279, 287, 318, 405, 406, 408, 412
Lachenmeyer, C. W. 38n2
Laetz, V. 233
LaGrassa, J. 143
Lana, R. E. 96n2
Lane, R. D. 367
Lang, P. J. 365
Langer, E. J. 142, 144
Langlais, B. T. 243
Laranjeira, R. 298
Larson, E. 174
Larson, R. 255
Lasswell, H. D. 285, 286, 288
Latane, B. 100, 113, 422
Lau, L. 293
Laurenceau, J.-P. 255
Lavrakas, P. J. 238, 240, 315, 316
Lawson, R. 126

- Lazarsfeld, P. H. 258
 Le, H. 362
 Le, T.- A. 322
 LeBel, E. P. 398
 LeCount, R. J. 317
 Ledermann, T. 374, 379
 Lee, A. 239
 Lee, Harper: *To Kill a Mockingbird* 282
 Lee, J. 145
 Lee, L. 317
 Lee, S. 293
 Lee, S.- G. 317
 Lefkowitz, M. 422
 Lehman, D. R. 254
 Leigh, B. C. 254
 Leinhardt, S. 382
 Leirer, V. O. 104, 355, 370n5
 Leitten, C. L. 364
 Lemeshow, S. 231
 Lepage, A. 101
 Lepper, M. R. 420, 421
 Lerner, D. 285
 Leventhal, H. 144, 259
 Levin, P. F. 141
 Levine, M. 139
 Levine, R. A. 156
 Levinson, D. J. 72
 Levy, P. S. 231
 Lewin, H. S. 287
 Lewis, A. N. 54, 56
 Lewis-Beck, M. S. 171, 172
 Li, X. 298
 Lickel, B. 362
 Lievens, F. 150
 Likert, R. 322–329
 Lin, A. 345
 Lin, S. S. 101
 Lincoln, Y. S. 256
 Lindquist, J. D. 251
 Lindsay, J. J. 138
 Lindsay, R. C. L. 124
 Link, M. W. 240
 Lipkus, I. M. 303
 Lippert, D. J. 287
 Liu, S. 214
 Livingston, M. 317
 Lo, L. L. 243
 Locander, W. 307
 Locascio, J. J. 211
 Logan, G. D. 351
 Lomax, R. G. 239
 Lord, C. G. 103
 Lord, R. 391n3
 Losch, M. 365
 Loving, T. J. 398
 Lu, F. 137
 Lu, J. 156
 Lubetsky, J. 380
 Luce, R. D. 114
 Luhmann, M. 364
 Lui, L. 115, 353
 Luke, J. V. 225
 Lumsdaine, A. A. 160
 Lundmark, M. 149
 Luria, Z. 326, 327
 Luskin, R. C. 317
 Lustig, C. 145
 Lutz, J. 145
 Luzier, J. L. 241
 Lynch, J. G. 204
 Lyons, J. B. 261
 Lyssenko, L. 156
 MacDougall, B. L. 315
 Macedo, J. 116
 MacKillop, J. 385
 MacLeod, C. 351
 Macrae, C. N. 360
 Madison, James 278
 Magelssen, M. 317
 Magilavy, L. J. 303
 Magura, S. 298
 Magyar, M. S. 112
 Mahoney, M. R. 243
 Maibach, E. 156
 Maislisch, L. 69
 Mak, A. 293
 Maki, J. E. 172
 Maldonado, J. H. 115
 Malhotra, N. 328
 Malone, P. S. 409
 Manicas, P. T. 4
 Mann, J. 116
 Mann, L. 111
 Manzi, J. M. 353
 Marcia, J. 292, 293
 Marcoulides, G. A. 41
 Markus, H. R. 409
 Marotzke, J. 115
 Marquis, K. H. 301, 307
 Marsh, H. W. 68
 Marshall, B. 235
 Martin, C. E. 297, 299, 305
 Martin, E. 310n4
 Martin, S. 241
 Martinez, L. S. 145
 Martinez-Rolan, X. 275
 Martinussen, M. 329
 Martsolf, G. R. 240
 Massey, D. S. 239
 Materstvedt, L. J. 317
 Mathewson, G. 36
 Mattavelli, G. 367
 Matter, C. F. 422
 Maurer, L. 287
 Mazza, G. L. 243
 McArthur, L. Z. 205, 351
 McAuliffe, W. E. 240
 McBride, C. M. 303
 McCain, John 105

456 Author Index

- McCall, C. 364
McCarthy, D. M. 298
McClendon, M. J. 330
McCredie, M. N. 146
McGatlin, K. 145
McGeeney, K. 239
McGhee 361
McGrath, J. E. 387, 389
McGraw, K. O. 146
McGuire, W. J. 6, 8, 92
McKenna, K. 391n2
McKenna, K. Y. A. 149
McNeish, D. 48
McNemar, Q. 160
Mcrae, C. N. 366, 368
McVey, D. 240
McWhorter, S. K. 82
Mearns, J. 69
Medin, A. C. 144
Mehan, H. 253
Mehl, M. R. 145
Meier, P. S. 239
Meisel, M. K. 385
Meisenhelder, H. M. 388, 389
Mellenbergh, G. J. 48, 56
Mellor, D. 429
Mendes, W. 362
Mende-Siedlecki, P. 398
Mendoza, J. L. 217
Meng, Y. 239
Merikangas, K. 143
Merrill, L. L. 82
Messe, L. A. 156, 387
Messick, D. M. 115, 116
Messick, S. 60, 72
Metzger, G. D. 239
Meyer, I. H. 240
Meyer, J. F. 69
Meyers, L. S. 328
Mezei, L. 422
Mickelson, K. D. 234
Midanik, L. T. 240
Middlemist, R. 422
Miles, A. 146
Miles, C. 307
Miles, J. 289
Miles, J. N. V. 52
Milgram, S. 111, 112, 124, 125, 131, 419, 420, 422
Milinski, M. 115
Miller, A. G. 111
Miller, C. N. 238
Miller, G. F. 298
Miller, J. 398
Miller, J. D. 385
Miller, N. 116, 365, 380
Miller, N. E. 6
Miller, P. V. 296, 298, 307
Miller, R. E. 380
Miller, S. E. 422
Mills, J. 35, 36, 258
Mills, J. F. 298
Mills, R. T. 144
Milne, A. B. 360
Milner, J. S. 82
Minder, C. 411
Mior, S. 241
Miron, D. 36
Mirsky, I. A. 380
Mischel, W. 355
Misra, R. K. 339
Mitchell, G. 362
Mitchell, R. C. 297
Mitsuhiro, S. S. 298
Mixon, D. 111
Moeck, E. K. 244n8
Moghaddam, F. M. 144
Mohammed, S. 389
Mokdad, A. H. 240
Molenaar, P. C. 214
Moll, A. 126
Monge, P. R. 385
Moniz-Cook, E. 361
Montaner, J. S. 235
Monteith, L. L. 238
Mook, D. G. 26, 131, 132
Moon, N. 240
Moore, B. 139
Moore, D. S. 412
Moosbrugger, H. 68
Mora Ringle, V. A. 241
Morain, S. R. 279
Moreland, R. L. 389
Moreno, J. L. 381
Moreno-Sanchez, R. del P. 115
Morey, L. C. 146
Morgan, R. D. 298
Moriarty, T. 249, 250, 270n1
Moroz, T. 367
Morris, C. A. 48
Morrison, D. M. 240
Morton, A. Q. 286
Mosco, B. 155
Moscovitch, M. 367
Moser, C. A. 228, 238
Mosteller, F. 278, 286, 338
Motevalian, A. 293
Mouton, J. S. 422
Mowrer, O. H. 6
Muchinsky, P. M. 49
Mueller, R. O. 48
Mukophadhyah, T. 149
Mulder, L. B. 115
Mullen, B. 390
Muller, G. 156
Murdock, G. P. 259
Murff, E. J. T. 112
Murphey, P. 381
Murrell, A. 116
Musch, J. 149

- Nakayama, K. 148
 Nam, C. S. 261
 Nanda, H. 48
 Napolitan, D. A. 204
 National Academies of Sciences, Engineering, and Medicine 150n1
 National Academy of Sciences 427
 Neale, M. C. 298
 Necka, E. A. 146
 Neely, J. H. 360
 Neidrich, R. W. 329
 Neilands, T. B. 242
 Nelson, C. 343
 Nelson, L. D. 428
 Nerlove, S. 343
 Neuendorf, K. A. 274, 280
 Newcomb, T. M. 380
 Nezlek, J. B. 255
 Ng, K. L. 293
 Nicasio, A. V. 322
 Niedenthal, P. M. 106, 345
 Ning, M. 145
 Nisbett, R. E. 107, 348, 409
 Nisbett, R. F. 258
 Nishisato, S. 340
 Noble, I. 240
 Noguchi, H. 328
 Noorbaloochi, S. 238
 Nordin, M. 241
 Norenzayan, A. 121, 409
 Norman, G. 328
 Norman, G. J. 146, 364
 Nortvedt, P. 317
 Nosek, B. 397, 429
 Nunnally, J. C. 44, 60, 318, 320
 Nussbeck, F. W. 68

 Oakes, L. 148
 Obama, Barack 105
 Ober, B. A. 344
 O'Connell, E. J. 380
 O'Connor, A. 156
 O'Connor, K. J. 368
 O'Connor, K. M. 389
 Ohtsubo, Y. 388
 Oksenberg, L. 296, 298, 307
 Oldendick, R. W. 297
 O'Leary, A. 156
 Olson, G. 389
 Olson, J. 351
 Olson, J. S. 389
 Olson, M. A. 307
 Open Science Collaborative 397
 Orbell, S. 102
 Orne, M. 124
 Ortmann, A. 418
 Orwin, R. 298
 Osborn, A. F. 390
 Osborn, L. 240
 Osborne, R. E. 361
 Osgood, C. E. 325–329, 326, 327
 Osterlund, C. 251
 Ostrom, T. M. 355, 370n5
 Oswald, F. L. 362
 Otto, M. 390

 Packer, D. J. 401
 Page, S. 422
 Paletz, S. 240
 Palmer, J. R. 167
 Paolacci, G. 241
 Papagno, C. 367
 Paris, C. 156
 Parker, S. 361
 Parker, S. D. 140
 Parker, S. E. 167
 Parkinson, S. 427
 Parks, M. J. 233
 Parrillo, V. N. 322
 Parry, H. J. 295
 Partin, M. R. 238
 Paskov, K. 145
 Patchett, E. 69
 Patel, N. 102
 Patel, Z. S. 241
 Patrick, M. E. 233
 Patten, E. 239
 Patterson, J. R. 71
 Patterson, V. 149
 Paul, B. Y. 365
 Payne, B. K. 361
 Pearce, L. 7
 Peebles, R. 288
 Pelham, B. W. 361
 Peluchetter, J. V. 390
 Pereira, A. 4
 Perrin, A. 239
 Perrotta, S. 143
 Peters, B. J. 255
 Peters, E. 255
 Peters, M. L. 144
 Petersen, M. M. 243
 Petri, J. M. 241
 Petty, R. E. 106, 130, 131, 317, 354, 364, 365
 Pfeiffermann, D. 239
 Phelps, E. A. 368
 Phillips, D. P. 289
 Piaget, Jean 8
 Picard, S. 298
 Pickett, C. L. 103
 Pierce, A. H. 124
 Pietromonaco, P. 357
 Pike, S. W. 25
 Piliavin, I. M. 139, 423
 Piliavin, J. A. 139, 423
 Pineiro-Otero, T. 275
 Piquero, A. R. 171
 Platania, E. 367
 Polifroni, M. A. 251
 Pollak, K. I. 303

458 Author Index

- Pomeroy, W. B. 297, 299, 305
Pool, I. de S. 285
Poole, K. T. 342
Poortinga, Y. H. 72
Popper, K. R. 4
Portenoy, R. K. 141
Portwood, A. C. 287
Poskanzer, D. C. 167, 168
Postmes, T. 428
Powell, M. C. 360
Prachthauser, M. 322
Prager, J. 398
Pratto, F. 351
Preacher, K. J. 205
Presser, S. 297, 317
Prislin, R. 102
Pryor, J. B. 355, 370n5
Przemieniecki, C. J. 251
Purshouse, R. 239
Pusch, M. D. 112
Puskar, C. 317

Qayad, M. 297
Qi, J. 235
Quadflieg, S. 366, 368
Quinn, R. P. 293
Quist, R. M. 143

Radgoodarzi, R. 293
Raffaelli, M. 156
Rahimi-Movaghar, A. 293
Rahn, W. M. 233
Raiffa, N. 114
Raifinan, J. 174
Rajaratnam, N. 48
Ram, N. 243
Ramirez, J. R. 171
Rasinski, K. 317
Rasinski, K. A. 317
Rasolofoarison, D. 351
Rast, D. E., III 143
Ratliff, E. A. 251
Rau, T. J. 82
Rauma, D. 172
Ray, M. L. 296
Raykov, T. 41
Raymond, P. 388
Reed, C. F. 126
Reed, F. A. 115
Rees, M. B. 73
Reichardt, C. S. 173
Reicher, S. 139
Reid, C. A. 54, 56
Reiman, E. 367
Reinero, D. 398
Reips, U. D. 146, 149
Reis, H. T. 255
Ren, Y. 388
Rensvold, R. B. 72
Renzaglia, G. A. 254

Resnick, J. H. 417
Resnick, S. G. 141
Resnicow, K. 322
Reynolds, M. L. 338
Rhodes, N. 92, 401
Rholes, W. S. 356, 357
Ric, F. 106
Richardson, D. 421
Richardson, J. T. 73
Richardson, M. W. 57n2
Richardson, S. A. 302, 305
Richter, L. 298
Riecken, H. W. 7, 252, 253
Rips, L. J. 317
Robert, R. S. 147
Roberts, R. R. 254
Robins, G. 384
Robinson, A. 174
Robinson, J. P. 254
Robinson, R. G. 367
Rodin, J. 139, 144, 423
Roemaker, D. L. 370n5
Rog, D. J. 26
Rogak, L. J. 243
Rogers, J. D. 240
Roitmann, E. 69
Rokeach, M. 422
Romney, A. K. 343
Room, R. 317
Rorer, L. G. 73
Rosenbaum, P. R. 162
Rosenberg, M. J. 111
Rosenberg, S. 343
Rosenthal, R. 23, 31, 122, 126, 127, 257, 399, 401, 404, 408, 410, 412, 413, 414, 419, 427
Rosenvinge, J. H. 329
Rosnow, R. L. 96n2, 412, 423, 427
Rosnow, R. W. 31, 122
Ross, L. 4, 409, 420, 421
Rossano, L. 240
Rossi, P. H. 294, 295
Roth, P. L. 242
Rotheram-Borus, M. J. 423
Rothschild, B. H. 111
Rothstein, H. R. 403, 405, 406, 415n4
Rottman, L. 377
Rovine, M. J. 214
Roy, A. K. 388
Rubin, D. B. 127, 162, 317, 401, 408, 413
Rubin, D. C. 322
Rucker, D. D. 102
Ruderman, A. 355
Rudert, S. C. 145
Rugg, D. 297
Russell, B. 258
Russell, R. I. 401
Ruth, T. J. 126

Sadler, M. S. 386
Saenz, D. S. 103

- Salas, E. 390
 Salmi, L. R. 241
 Salovey, P. 156
 Samek, A. 419
 Samuelson, K. W. 101
 Sanbonmatsu, D. M. 360
 Sanders, C. 297
 Sanford, R. N. 72
 Sanna, L. J. 4
 Sargis, E. 147
 Sawyer, J. 424
 Sawyer, S. 251
 Scanlon, D. P. 240
 Schachter, S. S. 7, 252, 253
 Schafer, J. L. 242
 Schamus, L. A. 240
 Schaps, E. 297, 422
 Schat, A. C. 287
 Schecter, H. 424
 Schenk, S. 288
 Scher, J. 144
 Schermelleh-Engel, K. 68
 Schillewaert, N. 73, 329
 Schimmack, U. 362
 Schippers, E. E. 7
 Schlenker, B. R. 120
 Schlerlis, W. 149
 Schmidt, B. 156
 Schmidt, F. L. 49
 Schmitt, M. 362
 Schmitt, N. 69
 Schneider, A. L. 238
 Schneider, J. 390
 Schneider, M. 411
 Schoen, L. 279
 Schofield, R. E. 240
 Schooler, N. 423
 Schorge, J. 167
 Schulenberg, J. E. 233
 Schultz, D. 111
 Schulzer, M. 126
 Schuman, H. 295, 297, 300, 304, 317
 Schwartz, G. 367
 Schwartz, G. K. 243
 Schwartz, H. 269
 Schwartz, J. 145, 361
 Schwartz, R. D. 143, 144, 422
 Schwartz, T. 417
 Schwartz-Bloom, R. D. 303
 Schwarz, N. 4, 70, 71, 146, 294, 330n1
 Sci, D. M. 241
 Scott, J. 384
 Scott, W. A. 341
 Scullen, S. E. 68
 Sears, D. O. 36, 121, 138, 427
 Sears, R. R. 6
 Sechrest, L. 143, 144, 422
 Secord, P. F. 4
 Seery, J. B. 342
 Seidell, J. C. 298
 Seo, J. 234
 Sequeira, L. 143
 Shadish, W. R. 35, 120, 156, 163
 Sharifi, V. 293
 Sharkness, J. 54, 57
 Sharma, S. 251
 Sharp, E. 259
 Shaver, P. R. 234
 Shaver, P. S. 234
 Shaw, J. S. 353
 Shearson, C. G. 317
 Sheffield, F. D. 160
 Shenaut, G. K. 344
 Shepard, R. N. 342, 343
 Sherman, J. 361
 Sherman, J. W. 398
 Shevlin, M. 52
 Shiffman, S. S. 338
 Shifman, L. 275
 Shinohara, S. 156
 Shiyko, M. P. 243
 Shrout, P. E. 255, 267
 Shulman, A. D. 420
 Shults, R. A. 297
 Sieber, J. E. 424
 Siebinga, V. Y. 241
 Siegel, J. T. 7, 102, 147, 163, 233
 Siegel, M. 174
 Siegel, P. M. 295
 Siegman, C. R. 73
 Siegrist-Dreier, S. 251
 Sigelman, L. 143
 Silverman, I. 420
 Silverstein, M. W. 241
 Silvia, P. J. 255
 Simmens, S. J. 240
 Simmons, A. 145
 Simmons, J. P. 428
 Simon, A. B. 238
 Simonsohn, U. 127, 428
 Simpson, D. D. 370n5
 Singer, E. 238, 239, 307
 Singer, J. 144
 Sivacek, J. M. 123
 Skitka, L. J. 147
 Skolnick, J. H. 422
 Slater, P. E. 387
 Sloan, L. R. 143
 Slovic, P. 4
 Smid, W. J. 7
 Smit, J. H. 7
 Smith, E. R. 25, 293
 Smith, G. D. 411
 Smith, G. T. 298
 Smith, H. S. 279
 Smith, M. L. 401, 402
 Smith, R. A. 137
 Smith, S. S. 421
 Smith, S. T. 112
 Smith, T. L. 333

- Snider, J. G. 325, 326, 327
 Snow, R. 31
 Snyder, J. 251
 Soderberg, C. K. 429
 Solomon, R. L. 81–82, 82, 85, **85**, 86
 Solomon-Moore, E. 145
 Sommerfeld, R. D. 115
 Sossi, V. 126
 Soto, C. J. 121
 Souchet, L. 303
 Spearman, C. 49
 Spears, R. 428
 Spence, J. R. 398
 Spencer, J. R. 140
 Spinner, B. 124
 Sprecher, S. 377
 Sproull, L. 389
 Sproull, L. S. 144
 Srivastava, S. 148, 149
 Srull, T. K. 357
 Stambaugh, R. J. 112
 Stancu, A. 155
 Stander, V. A. 82
 Standing, L. G. 317
 Stanley, D. J. 398
 Stanley, J. C. 19, 22, 27, 79, 120, 156, 160
 Stanley, T. D. 174
 Stasser, G. 117n3, 388
 Stein, Gertrude 289
 Steketee, G. 69
 Stemler, S. 274
 Stephan, D. 275
 Stevens, S. S. 12
 Stevenson, J. 238
 Stigglebout, A. M. 241
 Stigler, S. M. 415n2
 Stockham, N. 145
 Stocking, C. 307
 Stoessl, A. J. 126
 Stoll, C. 149
 Stone, P. J. 278, 285
 Stone, V. A. 287
 Stoops, W. W. 241
 Storreston, M. 389
 Stoto, M. 391n3
 Stouffer, S. A. 318
 Strack, F. 71, 397
 Strauss, A. L. 257
 Strauss, J. 143
 Streufert, S. 184
 Strickland, J. C. 241
 Striker, L. J. 421
 Stroebe, W. 307, 390, 397, 428
 Stroessner, S. J. 350
 Strohmetz, D. B. 124
 Strohsnitter, W. C. 167
 Stromberg, E. L. 254
 Stroop, J. R. 351–352, 361
 Studer, C. 144
 Stults, D. M. 69
 Stuss, D. 367
 Suchman, E. A. 320
 Suci, D. J. 325, 326, 327
 Sudman, S. 294, 296, 307, 317, 330n1
 Suedfeld, P. 184
 Sullivan, C. J. 171
 Suls, J. M. 96n2
 Sun, Y. 137
 Supphellen, M. 317
 Surgeon General's directives on human experimentation 417
 Swain, S. D. 329
 Swaminathan, H. 48
 Swann, W. B. 391n1
 Swanson, C. L. 295
 Szatmari, P. 143
 Tabata, L. 199
 Takane, Y. 69
 Takarangi, M. K. T. 244n8
 Takezawa, M. 388
 Tang, J. C. 7
 Tannenbaum, P. H. 325, 326, 327
 Tariq, Q. 145
 Taylor, S. E. 144, 350, 355
 Taylor-Vaisey, A. 241
 Tekman, H. G. 345
 ten Berge, J. M. F. 69
 ten Hove, D. 48
 Terry, R. 382
 Test, M. A. 422
 Tetlock, P. E. 184, 362
 Tew, M. D. 146
 Thanarajasingam, G. 243
 Thein, R. D. 361
 Theiner, G. 388
 Thistlethwaite, D. L. 173
 Thomann, S. 251
 Thomas, R. K. 328
 Thomas, S. 199
 Thomas, S. L. 199
 Thompson, C. L. 174
 Thompson, C. P. 370n5
 Thompson, S. 279, 287
 Thompson, S. C. 144
 Thomsen, C. J. 82
 Thorndike, R. L. 160, 415n2
 Thorne, A. 143
 Thorne, B. 253
 Thornhill, R. 298
 Thurstone, L. L. 318, 319, 320, **320**, 322, 323, 325, 330n2, 335, 345n2, 345n4
 Tian, G.- L. 388
 Timpano, K. R. 241
 Tindale, R. S. 388, 389
 Titus, L. 167
 Titus, W. 388
 Tolin, D. F. 69
 Tomaka, J. 364
 Tong, S. 389

- Torgerson, W. S. 340
 Tormala, Z. L. 102
 Torres, M. 116
 Tourangeau, R. 239, 317
 Trafton, J. A. 141
 Treger, S. 377
 Treiman, D. J. 295
 Trochim, W. 31
 Troisi, R. 167
 Trotter, R. T., II 385
 Trussel, N. 238
 Tsang, A. 293
 Tsuji, S. 148
 Tuchfarber, A. J. 297
 Tucker, W. 303
 Tulving, E. 367
 Tunnell, G. B. 249, 250
 Turkle, S. 149
 Turner, C. F. 310n4
 Turner, J. C. 34
 Tversky, A. 4, 338, 341
 Tweedie, R. 412

 Ueshima, N. 328
 Ulfelder, H. 167, 168
 Ulrich, M. 174
 Ulrich, R. 398
 Underwood, B. 139
 Uphoff, H. F. 34
 Upshaw, H. S. 320

 Vacanti-Shova, K. 423
 Vaisey, S. 146
 Valacich, J. S. 390
 Valadez, D. L. 240
 Valentine, J. C. 403, 413, 415n4
 Van Bavel, J. 398
 Van Bavel, J. J. 4, 368
 van de Ven, J. 419
 van der Ark, L. A. 48
 van der Zouwen, J. 294
 Van Dusen, R. A. 310n3
 van Heerden, J. 48
 van Herk, H. 72
 Van Horn, M. L. 322
 Van Lange, P. A. M. 115
 Vancleef, L. M. G. 144
 Vanman, E. J. 365
 Varma, M. 145
 Vazire, S. 148, 149
 Verhallen, T. M. M. 72
 Verplanken, B. 102
 Vevea, J. L. 406
 Viet, A. L. 298
 Vispoel, W. P. 48
 Visscher, T. L. S. 298
 Visser, P. S. 315, 316
 Vitaglione, G. D. 289
 Vivekananthan, P. S. 343
 Vogel, M. 244n8

 Vollrath, D. A. 388
 Voloudakis, M. 85
 von Brocke, J. 390
 von Hippel, W. 145, 251
 Vonderlin, R. 156
 Vorst, H. C. M. 56
 Voss, C. 145

 Wade, N. 428
 Wagge, J. 397
 Wagner, D. 71
 Wakita, T. 328
 Walker, D. S. 52
 Walker, M. R. 143
 Wall, D. P. 145
 Wallace, D. L. 278, 286
 Wallach, L. 120
 Wallach, M. A. 120
 Walsh, J. T. 293
 Walster, E. 205, 377, 420, 421
 Walther, J. B. 137, 145, 389, 423
 Wardenaar, K. J. 345
 Warner, R. M. 391n3
 Washington, P. 145
 Wasserman, S. 384
 Watt, S. 251, 252
 Watts, H. 156
 Weathers, D. 329
 Weathers, F. W. 241
 Weaver, C. N. 295
 Webb, E. J. 143, 144, 296, 422
 Webb, S. C. 320
 Webber, S. J. 124
 Wegener, D. T. 317
 Wegner, D. M. 388
 Weick, K. E. 250, 258, 265
 Weijters, B. 73, 329
 Weinberger, B. I. 156
 Weisenmuller, C. M. 241
 Weiss, N. H. 241
 Wellman, N. 145
 Wellman, N. S. 275
 Wells, G. L. 106
 Welsh, M. 48
 Wenzel, G. 139
 West, C. P. 238
 West, T. V. 4
 Westley, B. H. 287
 Weubben, P. L. 421
 Wheeler, L. 255
 Wheeler, S. R. 64
 Whelan, M. 401
 White, D. R. 259
 White, H. 383
 Whyte, W. F. 304
 Whyte, William Foote: *Street Corner Society* 304
 Wichers, M. 255
 Wiesenenthal, D. L. 420
 Wigman, J. T. W. 345
 Wilderquist, K. 156

462 Author Index

- Wilhelmy, R. A. 111
Wilke, H. 115
Williams, C. J. 360, 369
Williams, J. E. 146
Williams, K. D. 145, 146
Williams, W. 156
Williamson, J. 251
Williams-Piechota, P. 156
Willingham, D. T. 368
Willis, M. 251
Wilmer, J. B. 148
Wilson, D. 139
Wilson, G. D. 71
Wilson, J. L. 288
Wilson, R. 144, 145
Wilson, T. 25, 35, 132
Wilson, T. D. 107, 348
Winer, B. J. 96n4
Wing, C. 173
Winkielman, P. 106
Winocur, G. 367
Winograd, C. H. 126
Wish, M. 343
Witt, P. N. 126
Witte, T. K. 241
Wittich, C. M. 238
Wojcik, S. P. 146
Wondra, J. D. 377
Wong, H. 172
Wong, V. 173
Wood, E. 235
Wood, H. 253
Wood, S. J. 345
Wood, W. 92, 401
Wouabe, E. D. 173
Wright, C. 317
Wright, J. D. 294
Wyatt, R. C. 328
Wyatt, T. H. 308
Wyer, M. M. 260, 263, 265
Wyer, R. S., Jr. 357
Xu, C. 156
Xu, Y. 112
Yakobson, S. 286
Yang, P. 322
Yesavage, J. A. 401
Yeung, J. T. 317
Young, F. W. 338, 342
Yuan, Y. 389
Yule, G. U. 286
Yung, A. R. 345
Zadro, L. 145
Zander, A. F. 264
Zanna, M. 351
Zanna, M. P. 104
Zegiob, L. E. 254
Zemke, P. 115
Zhao, R. 298
Zhao, X. 204
Zhou, Y. 298
Zhu, P. 156
Zickar, M. J. 48
Zill, N. 310n3
Zillman, D. 7
Zimbardo, P. G. 112
Zimmerman, D. W. 48
Zlokovich, M. 397
Zoroufy, A. J. 243
Zuckerman, M. 213
Zuell, C. 283
Zwakhlen, S. 251

SUBJECT INDEX

Note: **Bold** page numbers refer to tables; *italic* page numbers refer to figures and page numbers followed by “n” denote endnotes.

- ACASI (audio computer-assisted self-interview software) 298
- acquiescence 72–73, 122, 141, 329; response bias 73, 329
- active role-playing simulations 111–112
- actor 6–7
- actor-partner interdependence model 379, 379
- agency 303–304
- alpha *see* coefficient alpha
- American Psychological Association’s Committee on Ethical Standards 417
- analogue experiments 112–116, 114; collective decisions 115–116; cooperation and competition in two-person experimental games 114–115; individuals 115
- anchor extremity 328
- anonymity 304
- answering machines and caller-ID 310n2
- applied research *vs.* basic research 26–27
- apprehensive participants 124
- archival data 174–175; *see also* content analysis
- artifact 157; threats to internal validity 70–73, 328–329
- ascendancy searching 402
- asymmetric dyad 377–379
- attention check 107
- attention, measures of 350–353; interference 351–352; processing time 352–353; visual 351
- attenuation 49, 52, 184, 324
- attitude scales *see* rating scales
- attrition *see* participant loss
- audio computer-assisted self-interview software (ACASI) 298
- autocorrelated errors 170, 171
- automatic evaluation 360, 369–370
- automaticity 361–362; cognitive busyness 361; response interference techniques 361–362
- autoregression path 212–213
- badness-of-fit indices 210
- balance theory 111
- bargaining games 112
- basic research 3–15; conceptual hypothesis 4–5; research hypothesis 5–6; science and daily life 4; *vs.* applied research 26–27
- between-dyad variables 375
- bias: experimenter (*see* experimenter expectancy); observer 34, 128, 254, 258; publication 410–412, 411; response 70, 329
- bidirectional causation 20
- binomial effect size display (BESD) 413, 413
- bipolar adjectives 325–327
- bivariate correlation and regression 181–189
- blind procedures 128–129; double-blind 126, 128–130
- blocked designs 91–92; *see also* factorial design
- blocked randomization 90–93
- blocking variable 90, 91, 91, 94, 95
- brain stimulation 366–367
- brainstorming groups 390
- brain wave measures 367
- California F Scale 72, 73
- cardiovascular indices 364–365
- case-control designs 167–168
- category systems 258–267; assessing use 263–264; Cohen’s kappa 265–266, **266**; construction 260–267; Fleiss’ kappa 266–267;

- intensity–extensity 260–261; interference 264–265; interrater reliability 265; number of categories 261–263, 262; time 260; unit of observation 259–260; variables 263, **263**; *see also* coding observations
- causal direction 20
- causation 19–22, 31, 164, 197; bidirectional 20; and experiments 21–22; moderators and mediators of relationships 20–21; noncausal covariation 20; third-variable 20; unidirectional 20; *vs.* covariation 20
- census 174, 222–226, 231, 232, 239, 243n1, 281, 282; *see also* U.S. Bureau of the Census; U.S. Census
- ceteris paribus 34
- change, assessment of 156–160; regression artifacts 156–157
- choice matrix 335, 336, 337, 339; pair comparison scaling 335–339, 336, **337**; prisoner's dilemma 114, 114–115
- classical measurement theory 40–49, 52, 54, 55, 57, 60, 68; assessing reliability 43–48; observed score 41–44, 47–49, 68, 159, 197, 391; random error 41; systematic error 42; true score 40, 41
- closed-ended questions 297, 300, 301, 315–316
- clustering 200, 214, 355, 356, 370n5
- cluster sampling 225, 226, 230–232, 231, 243; *see also* multistage sampling
- code of ethics 427; *see also* ethical standards; ethics
- coding observations 256–269; category systems 258–267; combining quantitative and qualitative 268–269; early data returns 257; interrater reliability 265–268; number of categories 261–263, 262; rating scales 267–268; structured–nonstructured dimension 256–258; units 259–260; variables 263, **263**; *see also* Cohen's kappa; Fleiss' kappa; intraclass correlation (ICC)
- coding unit 277–280; spatial and temporal 279; words and themes 278–279
- coding variable 264, 265, 267, 268, 277–280, 284, 285, 288–289
- coefficient alpha 45, 50, 323–325, 327
- coefficient of determination 159, 186, 192
- coefficient of reproducibility 321
- cognitive busyness 361
- cognitive dissonance 7
- Cohen's d 403–404
- Cohen's kappa 265–266, **266**, 285
- common events, observation of 7
- Common Rule 425
- communications and attitude change 102
- comparison group design 27, 160–161, 168, 171
- comparison time-series design 171–173
- complete counterbalancing 93, 94, **94**
- computer-assisted self-administered interviews (CAsAI) 241
- computer-mediated group interaction 389–390
- computer simulations 116n3
- concept priming 356–359; assessing awareness 358–359; subliminal 357–358; supraliminal 357
- conceptual hypothesis 5, 6
- conceptual replication 36, 132, 137–138, 144, 145, 398, 406
- concurrent validity 62, 63
- condition replicability 37
- confederate 6, 42, 100–103, 123, 125, 139, 145, 249, 250, 370, 376, 386, 399, 408, 409, 419, 420
- confidentiality, data 423–424
- confirmatory factor analysis 51, 51, 52, 206–207, 208, 208, 323, 324
- conflicting findings 6
- conflicting observations 6
- confound 27, 93, 99, 100, 137, 161–163, 165, 179, 193, 196, 197, 201, 289, 430; *see also* artifact
- constant 8
- construct validity 61, 66, 68, 132, 137, 138, 146, 399, 427; conceptual replication 132, 137–138, 144, 145; known groups method 64; meta-analysis 398–400; multitrait-multimethod matrix 66–67, 68–70
- contemporary measurement theory 48–57; factor analysis 50–52, 51; item response theory 52–57, 53, 56; latent factors and correction for unreliability 48–50
- content analysis 274–290; codebook create 276–277; coding systems 283–284; coding unit 277–280; deductive 275; effects analysis 289; identifying source 276; inductive 275; Internet 288; propaganda analysis 288–289; qualitative 275; quantitative 275; representative examples 285–289; sampling 280–283, 282; software 283–284; statistical analysis 284; stylistic analysis 289; summary of general paradigm 284–285; what was said 286–287; to whom was it said 287–288; who said it 285–286; *see also* coding unit; deductive content analysis; inductive content analysis
- content databases 282–283
- content validity 61–62, 65, **65**, 309, 319
- context unit 277–281
- control group 29, 31, **33**, 33–34, 79, 80, 81, 85, 95, 96n3; *see also* experimental design
- convenience sampling 234, 287
- convergent validity 63–66, **65**, 68, 147
- cooperation and competition in two-person experimental games 114–115
- correction for attenuation 49
- correlational analyses 190, 195–197, 289
- correlational design, advanced multivariate 199–217
- correlational research 179–198; causal analysis 19–22; curvilinear relations 183; nonlinear relationships 161; partial correlation 169–170; *see also* regression
- correlation coefficient 49, 159, 181, 184–186, 189, 377, 378, 404–405
- cost-benefit analysis 424, 425, 427
- counterbalanced ordering 94
- covariance, analysis 163
- cover story 102, 105, 113, 127, 354, 357, 359, 418, 420

- criterion 49, 189, 202, 217, 249, 308, 328, 398;
 - multiple regression 190–194, 199, 202; predictive validity 62–63, 65, 362, 370
- criterion validity 62–63, 65, **65**
- criterion variable 22, 24, 190, 197, 200, 202
- Cronbach's alpha 45, 49, 50, 57n2, 202, 314, 323, 329
- Cronbach's coefficient alpha 45
- cross-lagged panel models 211–214, *212*
- crowdsourcing 37, 145; websites 148
- data transitivity 337
- debriefing 99, 310n1, 420–422; funneled debriefing 358; process debriefing 421
- deception: in the laboratory 418–419; and participant well-being 419–420
- deductive content analysis 275, 290
- degree of residuals 189
- demand characteristics 128, 139, 142
- dependent variable 22, 25, 27–30, 79, 81, 83, 85–87, 89, 91–94, 99–102, 104–107, 111, 112, 131, 138, 139, 141, 143, 147, 161, 179, 193, 249, 348, 353, 356, 360, 399, 401, 403, 404, 412, 429; assessing in field settings 142, 249, 348; measurement validity 144, 161; memory 353
- descendancy searching 402
- diaries: structures 254, 255; use in research 263
- difficulty parameter 54
- direct effects 204–206, 209
- discriminant validity 64–67, **65**, 69; combining with convergent validity 66–67
- discrimination parameter 54–55, 55
- disguised field experiment 429–430
- disproportionate stratified random sampling 229–230
- disturbance terms 209
- double-barreled question 295, 315
- double-blind procedures 126, 128–130
- dyadic measures 374–391; group-level variables 376–377; measuring reciprocity or mutuality 377–379
- dyadic variables 374–377, **375**
- early data returns 257
- ecological validity 130–132
- EEG (electroencephalography) 367
- effect size 24, 98, 116n1, 181, 185, 192, 398, 400–406, 403, 409–414; heterogeneity 406–408, *407*
- electroencephalography (EEG) 367
- endogenous variables 202–203, 205, 206, 212
- entry 252–253, 302–303; after entry 303–304
- environmental manipulation 100–101, 147
- epsem designs 227, 229
- epsem method 228
- equal appearing intervals 318–320, 322
- error rate 34; Type I 23, 24, 42, 43, 397, 428; Type II 24, 42, 43, 92, 98; *see also* bias; random error
- ESP *see* extrasensory perception (ESP)
- ethical standards 417, 425, 427
- ethics: codes of 427; data reporting 427–429;
 - deception and participant well-being 419–420; deception in the laboratory 418–419; issues related to the products of scientific research 429–430; methodology as 427; research practices 417–424; *see also* informed consent
- evaluation research 156, 170
- event-related potential (ERP) 366, **366**, 367
- exact replication 137, 138, 397–398, 406
- excitation transfer 7
- exogenous variable 202, 205, 206
- expectancy effects 126, 127, 130, 257; early data returns 257
- experience sampling method (ESM) 255
- experimental control *vs.* statistical control 193–194
- experimental demand 127–128
- experimental design 79–96; basic variations 79–82; blocked designs 90–93, *91*; expanding the number of experimental treatments 82–90; factorial designs 83–85, **83–85**; interaction effects 86–90; posttest-only control group 81; pretest–posttest control group 29, 30, 31, 32, 79–81, 80, 93, 155, *166*, 173; repeated-measures designs and counterbalancing 93–95; Solomon four-group design 81–82, 82, 85, 85
- experimental manipulations 100–106; environmental 100–101; instructional 103–104; priming 104–106; social 102–103; stimulus 101–102
- experimental methods 18, 22, 38n2, 80, 136, 175, 193, 217, 269, 348, 401; *see also* quasi-experimental methods
- experimental realism 109, 110, 113, 116, 125
- experimenter expectancy 133n2, 257, 305; and bias 125–130
- experiments outside the laboratory 136–150;
 - field experiment 138–144; Internet as a site for experimental research 144–150; research settings and issues of validity 137–138
- exploratory factor analysis (EFA) 50–51, 324
- external validity: basics issues of 35–37; definition 22; and internal validity 18–38; of laboratory experiments 120–133; *see also* ecological validity; generalizability; relevance; robustness
- extrasensory perception (ESP) 34
- extreme-response sets 72
- face-to-face interviews 221, 232, 238, 240–241, 243
- face validity 61
- facial action coding system (FACS) 365
- facial electromyograph 365
- factor analysis 48, 50–52, 54, 57, 62, 323–325, 343;
 - confirmatory 51, *51*, 52, 206–207, *208*, 323, 324; exploratory 50–51, 324
- factorial design 83–87, 90–92, 95, 96n4, 108;
 - creating the design 83–84; example 85; factor independence 84; mixed 91; treatment interaction effects 85
- faithful participants 124
- false consensus effect 4
- feasibility studies 167

466 Subject Index

- Federalist Papers* 278, 286
field experiments 138–144; application 144; control over independent variable 138–140; dependent variables 142; random assignment 140–142; selection of participants 138; unobtrusive measures 142–144
field research: *vs.* laboratory research 25–26; *see also* field experiments; observational methods
file drawer problem 410
fit indices 210–211
fixed-effect model 405, 406
Fleiss' kappa 266–268, 285
focus groups 274, 308–309
foot-in-the-door technique 141, 303
forced choice measure 317
free recall 354
functional magnetic resonance imaging (fMRI) **366**, 366–368
funnel plot 411

gaining entry 252–253
galvanic skin response (GSR) 364
gatekeeper 302–303
generalizability 35–37; across participants 121–125; across people 36–37; across settings 36–37; of operationalization 35–36; *see also* external validity
generalizability theory 48
general sampling weight 239
generations design 386
going native 253–255
goodness-of-fit indices 14, 210
good participants 124
G*Power 98
group interview 308–309
group level analysis 385–387; nested design 199, 200, 202, 270n3, 385–386
group level variables 376–377, 387, 391n2
group performance 387, 390
group process measures 387–390
group variables 374, 376
growth curve model 199, 214–216, 214–217
guessing parameter 55, 56
Guttman's scalogram 320–322, 321

hang together 44, 45, 50–52, 208
HARKing 428, 429
heterogeneity 46, 232, 406–409
hidden profile technique 388
hidden third factor 196, 196–197
hierarchical linear modeling 199
hierarchically nested design 385
hypotheses: conceptual hypothesis 5–6; research hypothesis 5, 8, 112, 122, 124, 128, 144, 149–150, 400, 418
hypothesis generation 6, 7, 129, 248, 252
hypothetico-deductive model 6

implicit association test (IAT) 361–362, 363
implicit measures 360, 362, 369, 370
implicit responses 350, 369
improving nonequivalent groups designs 165–167
independent variable 22, 27–30, 35, 82–87, **83**, **84**, 89–91, 93–95, 99–101, 106–109, 113, 137, 138, 155, 190, 193, 356, 399, 400, 401, 403, 412, 428; control over 138–140
indirect effects 204–206, 209; *see also* mediator variables
indirect measures 142, 143, 349–350; *see also* sentence completion task; thematic apperception test
individual performance 390
inductive content analysis 275, 290
inference 264
information processing measures 350–356; attention 350–353; memory 353–356; processing time 352–353
informed consent 99, 124, 417–419, 423
institutional review boards (IRB) 99, 147, 425–427, 426, 431n3
instructional manipulation 103, 104, 108
instrumentation effects 29, 34
instrumentation error 28
interaction effects 36, 86–90; forms 88–89; moderator variables 89–90; treatment 85
interaction process analysis (IPA) 261, 263, 269, 387
interdependence model 379, 379
internal consistency 44–46; *see also* Cronbach's alpha; split-half reliability
internal validity: basic issues 27–29; causation 19–20; and experiments 21–22; definition 22; distinguishing 22–27; and external validity 18–38; participant loss 32; random assignment and experimental control 29–31; reprise 32–33; *see also* validity
Internet: conducting experiments online 145–146; experimental research 144–150; methodological concerns 146–147; misconceptions about research 148–150; online studies 147–148; privacy 423; research 148–150; research on social network activity 145
Internet surveys 241
interrupted time-series design 169, 171
interval scale 12, 337, 339, 340
interviewer characteristics 304–305
interviews 292–310; administration 292–293; after entry 303–304; attitudes, intentions, and behaviors 296–297; closed-ended questions 300–301; conducting 301–302; demographic information 295–296; direction 306–307; face-to-face 240–241; group 308–309; informal tactics 305–308; leading questions **305**, 305–306; obtaining entry 302–303; open-ended questions 300; question content 294–298; reconstruction from memory 296; semi-structured 299; sensitive beliefs and behaviors 297–298; structure 298–301; structured 299–301; telephone *vs.* face-to-face 240–241, 292–293; unstructured 298;

- verbal and nonverbal reinforcement 307–308;
see also closed-ended questions; focus groups;
gatekeeper; group interview; leading question;
open-ended questions; question constraint;
question content; question specificity; self-
stimulation effects
- intraclass correlation (ICC) 165, 200–201, 267, 268,
268, 285, 378, 379, 381
- involuntary participants 122–123, 124
- IRB *see* institutional review boards (IRB)
- item response theory 52–57, *53*, *56*; parameters of
item characteristic curves 54–55; simultaneously
estimating all three parameters 55–57
- item-reversal 72
- judgment experiments 101–102
- kappa *see* Cohen's kappa
- known groups method 64
- laboratory environment, set up 99–100
- laboratory experiments, constructing 97–117;
engagement 109–110; manipulation and
attention checks 106–107; participant pool
identification 97–98; prepare materials 99;
randomization procedures 107–109; realism
109–110; role-playing simulations and analogue
experiments 110–116; sample size 98; set up
laboratory environment 99–100; submit plan to
IRB 99; types of experimental manipulations
100–106
- laboratory research 7, 25, 26, 34, 110, 138, 139, 146,
369, 418; neuroimaging measurements 369; *vs.*
field research 25–26
- language, difficulty 71–72
- latency *see* reaction time measures
- latent factor 48–54, *51*, *51*, 56, 57, 60, 68, 199, *203*,
207–210
- latent structural equation model 199, 202, 206–210,
207–209
- latent variables 277, 283, 289, 290
- law of large numbers 30, 90
- leading question **305**, 305–306
- lexical decision task 359–360
- Lexis-Nexis 283
- Likert scale 322–325, **326**, 326–328
- linear trend 214–216, *215*
- literature: database 402; understand 400
- longitudinal research 174–175, 424
- main effects 84–89, 91, 164
- manifest variable 277
- manipulation check 106–107
- matching: post hoc 160–161; propensity score
161–163; *see also* regression artifacts
- maturation effects 27, 28
- measurement component 206
- measurement construct validity 61
- measurement error 41, 45, 48–49, 51, 52, 68, 69, 184,
197, *203*, 206–210, 223, 314, 323, 333, 336, 345n1
- measurement model 325
- measurement reliability 40–57, 144, 146, 184,
210; classical measurement theory 40–48;
contemporary measurement theory 48–57
- measurement scale 12
- measurement validity 60–73; assessment 146–147;
threats 70–73, 329; types 61–70
- median split 180
- mediation 194–195; full 194; partial 194
- mediators 20–21, 194, 195; multiple 205
- mediator variables 21
- meditational processes 204
- memory measures 356; recall 353–355; recognition
355–356
- meta-analysis 398–400; interpreting 406–414, *407*,
409, *411*, *413*; locating primary studies 402–403;
stages 400–406; study inclusion and exclusion
criteria 400–402; understand the literature 400
- meta-moderators 408–410, **409**
- Minnesota Multiphasic Personality Inventory
(MMPI) 61, 73
- missing at random (MAR) 242, 243
- missing completely at random (MCAR) 242, 243
- missing data 241–243; missing at random (MAR)
242, 243; missing completely at random
(MCAR) 242, 243; not missing at random
(NMAR) 242–243; possible solutions to 243
- mixed design mixed factorial design 91; *see also*
blocked designs
- moderators 20–21, 130, 308–309, 408–410;
demographic 408; study-level 408–409;
theoretical 408
- moderator variables 21, 92, 308–309, 401, **409**;
heterogeneity 407–409; interaction effects
89–90, *90*
- monadic variables 374
- monotone items 320
- mood effects 70
- mortality 28, 32
- MTMM *see* multitrait-multimethod matrix
(MTMM)
- multidimensional scaling 340–345, **341**, *342*, *344*;
unfolding **341**, 341–342, *342*
- multilevel modeling 199–202, *200*, *201*, 385
- multilevel regression modeling 199
- multiple mediation 205–206, *206*
- multiple operationalization 10–12, 14, 66
- multiple R^2 192–193
- multiple regression 23, 160, 190–194, *195*, 199–200,
202–204, 216–217, 284, 379; experimental
control *vs.* statistical control 193–194; multiple
 R^2 192–193; standardized regression equation
191–192; unstandardized regression equation
190–191
- multistage sampling 225, 230–232, *231*, 243; panel
surveys 233; stratified 228–230, 232, 235
- multitrait-multimethod matrix (MTMM) 66–68;
evaluating results 68–70
- multivariate correlational designs 199–217; cross-
lagged panel models 211–214, *212*; growth curve

468 Subject Index

- models 214–216, *214–216*; multilevel models 199–202, *200, 201*; structural equation models 202–211
- mundane realism 109–110, 125, 132
- narrative review 398, 413; *vs.* vote counting *vs.* meta-analysis 413–414
- National Merit Scholarships 18
- National Security Agency 423
- naturalism 248–251; behavior 249; event 249–250; setting 250–251
- naturalness 249
- natural randomization 140
- negative participants 124
- nested design 200, 214, 270n3, 385–386
- neuroimaging 367–369
- nominal groups 390
- nominal scale 12, *13*
- noncausal covariation 20
- nonequivalent groups designs 164–167; improvement 165–167; primitive 164–165
- nonexperimental methods 18, 19, 193
- nonexperimental research 179–198; multiple regression 190–194; uses and misuses of correlational analysis 195–197, *196, 197*
- nonindependence, data 385
- nonlaboratory research, ethical issues 422–423
- nonlinear relationships *183, 183–184, 186*
- nonmonotone items 318
- nonprobability sampling 226
- non-random sampling 226, 233–234; convenience 234; quota 235; snowball 234–235
- nonresponse, surveys 237–239
- nonvoluntary participants 123–124
- “not-at-homes” (NAHs) 238
- not missing at random (NMAR) 242–243
- null hypothesis 23, 24, 38n4, 92, 147
- null hypothesis significance testing (NHST) 98
- objectivity 3
- observational methods 248–270; coding 256–269, *257, 262, 263, 266, 268*; concealed 254; naturalism 248–251; nonparticipatory distinction 253–256; participatory 251–253
- observed score 41–44, 47–49, 68, 159, 197, 391
- observer bias 34, 128
- odds ratio 403, 404
- online experiments 145–146
- online interviews 293
- online studies 147–150
- open-ended questions 297, 300, 315–316, 353
- operational definition 9
- operationalization 8–12, 15, 138, 140, 323, 399; definitions 8; generalizability 35–36; imperfections 9–10; measurement 40–57; multiple 10–12, *11, 14, 66, 323*; variable 8, 101
- order effects *see* question ordering; repeated measures
- ordinal scale 12, *13, 337*
- Osgood’s semantic differential 325–327, **326, 327**
- oversampling 229
- overt behavior 348, 370
- pair comparison scaling 335–339, *336, 337, 339, 340, 343*
- panel survey 233
- paradoxes 7
- parafoveal processing 358
- parallel forms reliability 47
- partial counterbalancing 94
- participant loss 32
- participant roles 124–125; apprehensive participants 124; faithful participants 124; good participants 124; negative participants 124
- participants: awareness 121–124; restriction of populations 121; voluntary 122
- participant selection 32, **33, 163**; randomization procedures 107–109
- participant well-being 419–420
- participatory observation 251–256; gaining entry 252–253; going native 253; partial 251–253; when and why 255–256
- partner 374
- passive role-playing simulations 111
- path models 202–206, *203, 204, 206*; mediation 204–205; multiple mediation 205–206, *206*
- Pearson correlation 181, 184–186, *187*
- Pew Research Center 424; Project for Excellence in Journalism 287
- phantom construct 11
- phases of research *see* dependent variable; independent variable
- phi correlation (ϕ) 185, *187*
- placebo effect 126
- point-biserial correlation 185, 186, *187*
- population parameters 224, 237
- Portrait of the Artist as a Young Man* 286
- post hoc matching 160–161
- posttest-only control group 79
- potency 326
- power: analysis 98; statistical 24, 98, 174, 233, 235, 280, 297, 399, 412
- practice effect 29
- precision and sampling error 222–225, *223, 224*
- predictive error 202, 203, 209, 212, 217
- predictive validity 62–63, 65, 362, 370
- predictor variable 21, 22, 36, 188, 190, 191, 197, 202, 217
- President’s Panel on Privacy and Behavioral Research 417
- pretest-posttest control group design 29, *30, 31, 32, 79–81, 80, 93, 155, 166, 173*
- pretest sensitization 80, 81, 85, 86, 93, 96n2
- primary study 397, 399, 402, 403–405, **404, 407–409, 411, 412**
- priming manipulation 100, 104–106, 358, 362
- priming techniques 356–362; assessing awareness 358; concept 356–359; sequential 359–361; subliminal 357–358; supraliminal 357

- prisoner's dilemma 114
- Prisoner's Dilemma Game 114, *114*, 115
- privacy: Internet 423; invasion 122, 422
- probability sampling 226, 233
- processing time 352–353
- program 38n1
- program evaluator 161
- propaganda analysis 288–289
- propensity score 161–163, *162*
- propensity score matching 162–163
- proportionate stratified random sampling 22, 229
- Psych Experiments 146
- Psychological Science 428–429
- psychophysiological measures: cardiovascular
 - indices 364–365; facial electromyograph 365;
 - neuroimaging 367–369
- publication bias 410–412, *411*
- public behavior 422
- quadratic trend 215
- qualitative content analysis 275
- quantitative content analysis 275
- quasi-experimental designs 163–174; case-control
 - 167–168; nonequivalent groups 164–167;
 - regression discontinuity 172–174; time-series 168–171
- quasi-experimental methods 18, 80, 156–160, **158**, 175; regression artifact and reliability 157–160, **158**; regression artifacts and assessment of change 156–160; statistical control of regression artifacts 160
- quasi-experiments: comparison group 27, 160–161, 168, 171; pretest–posttest 27, 28, 29, *30*, 79–81, 80, 93, 173; regression discontinuity 172–174, *173*
- question constraint 297, 300
- question content 294–298, *295*
- questionnaires 314–330; dropout and no-opinion response format 317; open- and closed-ended questions 315–316; question order 316–317; question wording 315; *see also* double-barreled question; Guttman's scalogram; rating scales; Thurstone scales
- question ordering 316–317
- question specificity 297
- quota sampling 226, 235
- random assignment 18, 34, 37, 79, 81, 84, 90–92, 107, 108, 109, 139, 155, 156, 164, 193, 221, 222; experimental control 29–31, *30*; field settings 140–142
- random digit dialing 226, 239–240
- random-effects model 405, **405**, 406–408, 412
- random error 40, 41, 42, 43, 49, 52, **158**, 159, 203
- randomization, blocked 90–93
- random numbers 223, 227, 228, 239; table 107, 223, 227, 244n4
- random sampling 222, 226, 230; simple 226–227, 228, 232, 235, 243n3; stratified 228–230, 235
- random selection 37, 233
- rank order scaling 339
- Rasch model 54
- rating scales 72, 259, 267–268, **268**, 314, 315, 333, 335, 349, 357, 375; constructing 318–329, **321**, **326**, *327*
- ratio scale 12, 340
- reaction time measures 360
- realism: experimental 109, 113, 116, 125; mundane 109–110, 125, 132
- recall measures 353–354, 370n2
- reciprocity 202, 203–206, *204*, 377–379, **378**, *379*, 382, 390–391
- recognition measure 353, 355–356
- regression: multiple 190–194, *195*, 199, 200, 202, 203, 204, 216–217, 284, 379; with single predictor 186–189; statistical 28; to test mediation 194–195
- regression artifacts: assessment of change 156–160, **158**; mean and matching 160–161; reliability 157; statistical control 160
- regression discontinuity (RD) designs 172–174, *173*
- regression equation 187
- regression line 187, *188*, 189, 190
- reinforcement, verbal and nonverbal 307–308
- relations 8
- relevance 132
- reliability: assessing 43–48; Cronbach's coefficient alpha 45, 48, 49, 50, 57n2, 314, 323, 329; definition 40; internal consistency 44–46; regression artifacts 157–160, **158**; split-half 44–45; temporal stability 46–47
- repeated measures: definition 93; designs and counterbalancing 93–95, **94**
- research hypothesis 5, 7, 122, 124, 128, 129, 144, 149, 167, 419
- research question 6, 179, 274, 276, 286, 366, 377, 398, 399, 419, 420, 429; deception 419–420
- research settings and issues of validity 137–138
- response formats 327–328; anchor extremity 328; bipolar *vs.* unipolar response ratings 328; number of response options 328; use of reversed items 328–329
- restriction of range 184
- retrieval bias 402
- robustness 36, 37, 130–131, 132
- role-playing simulations 110; active 111–112; passive 111; as research 116–117
- round robin designs 379–381, **381**
- sample size 98, 146, 224, 235–237
- sampling 222–226, 280–283, 282; cluster 226, 230, 231, 232; Internet surveys 241; multistage 230–232, *231*; random 222, 226, 230; simple random 226–227, 228, 232, 236, 243n3; source and content 281–282, 282; stratified 228–230, 232, 235; survey basics 222; two-stage 232–233; use of archives 282–283; *see also* sample size
- sampling error 23, 222–225, *223*, **224**, 406
- sampling fraction 224, 226–229, 235–236, 243n2
- sampling frame 223, 225–230, 232, 233, 237, 239, 241

470 Subject Index

- sampling weight 238, 239
- scaling: Guttman's scalogram analysis 320–322, **321**;
 - Likert's method of summated ratings 322–324;
 - method of rank order 339–340; multidimensional 340–345, **341**, 342, 344; Osgood's semantic differential 325–327, **326**, 327; pair comparison 335–339, 336, **337**; stimuli 333–345; Thurstone's method of equal-appearing intervals 318–320, **320**, 325, 330n2, 345n2; unfolding technique **341**, 341–342, 342
- scalogram analysis 322
- scatter plots 181, 182, 183, 183, 186, 187, 188, 188, 189, 411
- scientific method 3–4, 21, 348, 397
- scrambled sentence task 357
- selection 28, 30, **33**
- selection-based interaction 28
- selection bias 31, 157, 160, 397
- selection error 28
- self-consciousness effects 121
- self-report 66, 67, 70, 71, 100, 105, 107, 142, 143, 145, 254, 292, 298, 314, 316, 330, 348, 349, 360, 362, 364, 365, 369, 376, 384, 408
- self-selection 30, 31, 139, 140
- self-stimulation effects 300
- semantic differential 325–327, **326**, 327, 399
- semi-structured interviews 298, 299
- sentence completion task 349
- sequential priming 359–361; automatic evaluation 360; issues related to use of reaction time measures 360–361; lexical decision task 359–360
- sham operation 34
- shrinkage 157, 324
- simple random sampling 226–227, 228, 232, 235, 243n3
- simulation research 111, 112, 116n3
- snowball sampling 226, 234–235
- social cognition 350, 356, 360, 369, 370n5, 374; *see also* implicit measures
- social desirability 71
- social manipulation 102–103, 103
- social network analysis 145, 384–385
- social psychophysics *see* stimulus scaling
- social psychophysiology 362–369; brain activity 366–369; cardiovascular indices 364–365; facial measures of affective states 365–366
- social relations model 380
- social responsibility and ethics in social research 417–431; ethical issues related to the products of scientific research 429–430; ethics of data reporting 427–429; ethics of research practices 417–424; regulatory context of research involving human participants 424–427, 426
- social simulations: and analogue experiments 112–116; role-playing 110–112
- sociogram 383
- sociomatrix **382**, 382–383
- sociometric analysis 381–383
- sociometry 381
- Solomon four-group design 81, 81–82, 82; example 85
- spatial unit 279
- split-half reliability 44–45
- sponsorship of research 429
- spurious 20
- spurious relationship 20, 157
- stages of research 400–406
- standard deviations (SDs) 159
- standard error 223, 224, 225, 236, 243n2
- standardized regression equation 191–192
- startle eyeblink reflex 365–366
- statistical conclusion validity 23–25, 24
- statistical power 24, 92, 98, 174, 233, 235, 280, 297, 399, 412
- statistical regression 28; *see also* artifact
- statistical significance 23–24, 38n3, 41, 43, 98, 211, 398, 399, 400, 411, 412, 414
- stimulus manipulation 100, 101–102, 108
- stimulus scaling 333–345; techniques 335–339
- stratified sampling 228–230, 232, 235
- Stroop effect 351–352, 361
- structural component 206
- structural equation models 202–211; fit 210–211; fit indices 210–211; latent structural equation models 206–210, 207–209; path model 202–206, 203, 204, 206; path models 202–206, 203, 204, 206
- structured interview 299–300
- stylistic analysis 289
- subliminal priming 357–358
- subscales 44, 50, 54, 73, 324
- summary effect **405**, 405–406, 407, 408, 409, 410, 411, 412–413, **413**; fixed-effect model 405, **405**, 406; random-effects model 405, **405**, 406, 411, 412; unweighted model 405, **405**
- summated ratings 322–324
- survey research: answering machines and caller-ID 243; Internet 241; missing data 241–243; non-random sampling 233–234; random digit dialing 239–240; random sampling 226–233; selection *vs.* assignment 221–222; telephone *vs.* face-to-face interviews 240–241; types 239–241; *see also* interviews; sampling
- symmetric dyad 377, 378, 379
- synchronous correlation 211–212
- synthesis 144, 397, 399, 401, 402, 406, 410, 415n2, 415n5; quantitative (*see* meta-analysis)
- systematic error 40, 41, 42, 43, 44, 52, 159
- systematic sampling 227–228
- telephone interviews 240, 292–293
- temporal stability 46; parallel forms reliability 47; test-retest reliability 46, 67, 157, **158**, 159, 160, 285
- temporal unit 260
- test-retest reliability 46, 67, 157, **158**, 159, 160, 285
- thematic apperception test 349
- theoretical meta-moderators 408

- theory 4–6, 5; generating a hypothesis 6–8; operation to measurement scales 12–13, 13; role of in research 13–14; *see also* operationalization
- Theory of Reasoned Action 207
- threats *see* artifact, threats to internal validity; measurement validity, threats; validity, threats to experiments
- Thurstone scales 318–320, **320**, 322, 325, 330n2; reliability 319–320, **320**
- time and technical expertise 350
- time-series design 168–172, 169, 170, 172; archival data; autocorrelated errors 174–175; comparison 171–172, 172; interrupted 169–170, 170; prewhitening 171
- transcranial magnetic stimulation 366
- translation 4, 8–9, 40, 424
- treatment *see* manipulation check
- triangulation 10, 11, 11, 66; *see also* multiple operationalization
- true experiment 30, 79, 93, 107, 155, 156, 163, 164, 172, 174
- true score 40–43, **42**, 44, 48, 49, 68, 157, 158, 159, 160, 197, 323
- two-phase sampling 232–233, 244n6
- two-stage sampling 231
- Type I error 23, 24, 42, 397, 428
- Type II error 24, 42, 43, 92, 98
- unfolding technique **341**, 341–342, 342
- unidirectional causation 20
- unintentional processes 34, 38n5, 80, 127, 252, 369
- unit of analysis 87, 96n1, 260, 267, 384, 385, 403; nested designs 385–386
- unit of observation 259–260
- unmeasured variables 30, 242
- unobtrusive measures 142–144; ethics 422
- unstandardized regression equation 190–191
- unstructured interviews 298
- unweighted model 405, **405**
- U.S. Bureau of the Census 232, 424
- U.S. Census 239, 243n1
- validity: concurrent 62, 63; construct 63, 64, 65, 66, 68, 132, 136, 137, 138, 146, 399, 427; content 61–62, **65**, 319; convergent 63–64, **65**, 66, 68, 147, 369; criterion 62–63, **65**; definition 40; ecological 131–132; external 18–38; face 61; internal 18–38; measurement 60–73, 144, 146–147, 310, 329; operationalization, predictive 62–63, 65, 362, 370; statistical conclusion 23–25, 24; threat to experiments 70–73, 329; *see also* construct validity
- variable 8; between-dyad variables 375; blocking 90, 91, 91, 94; category systems 263; coding observations 260–261; criterion 22, 24, 190, 197, 200, 202; definition 8; dependent 22, 25, 27, 29, 30, 79, 81, 83, 84, 85, 86, 87, 89, 91, 92, 93, 94, 99, 100, 101, 102, 104, 105, 106, 107, 111, 112, 113, 131, 138, 139, 141, 142, 143, 144, 147, 161, 179, 193, 249, 348, 353, 356, 360, 399, 400, 401, 403, 404, 412; group level 83–85, **85**; independent variable 22, 27, 28, 29, 30, 35, 82, 83, **83**, 84, **84**, 85, 86, 87, 89, 90, 91, 93, 94, 95, 99, 100, 101, 106, 107, 108, 109, 113, 137, 138–140, 155, 190, 193, 356, 399, 400, 401, 403, 412; latent 277; manifest 277; moderator 21, 89, 92, 308–309; monadic 374
- verification 3, 4, 38n2, 107, 258
- voluntary participants 122, 124
- vote counting 413–414
- within-dyad variables 375
- within-subject designs *see* repeated measures
- World Wide Web *see* Internet
- zero correlation 181, 182, 186
- Z score 191, 192



Taylor & Francis Group
an informa business

Taylor & Francis eBooks

www.taylorfrancis.com

A single destination for eBooks from Taylor & Francis with increased functionality and an improved user experience to meet the needs of our customers.

90,000+ eBooks of award-winning academic content in Humanities, Social Science, Science, Technology, Engineering, and Medical written by a global network of editors and authors.

TAYLOR & FRANCIS EBOOKS OFFERS:

A streamlined experience for our library customers

A single point of discovery for all of our eBook content

Improved search and discovery of content at both book and chapter level

REQUEST A FREE TRIAL

support@taylorfrancis.com

 **Routledge**
Taylor & Francis Group

 **CRC Press**
Taylor & Francis Group