



Third Canadian Edition

METHODS IN BEHAVIOURAL RESEARCH



PAUL C. COZBY RAYMOND A. MAR CATHERINE D. RAWN

Methods in Behavioural Research

Third Canadian Edition

Paul C. Cozby, Ph.D.

California State University, Fullerton

Raymond A. Mar, Ph.D.

York University

Catherine D. Rawn, Ph.D.

University of British Columbia





Methods in Behavioural Research Third Canadian Edition

Copyright © 2020, 2016, 2012 by McGraw-Hill Ryerson Limited.
Copyright 2012, 2009, 2007, 2004, 2001, 1997, 1993, 1989, 1985, 1981,
1977 by McGraw-Hill Education LLC. All rights reserved. No part of this
publication may be reproduced or transmitted in any form or by any means,
or stored in a data base or retrieval system, without the prior written
permission of McGraw-Hill Ryerson Limited, or in the case of
photocopying or other reprographic copying, a license from The Canadian
Copyright Licensing Agency (Access Copyright). For an Access Copyright
license, visit www.accesscopyright.ca or call toll free to 1-800-893-5777.

The Internet addresses listed in the text were accurate at the time of
publication. The inclusion of a Web site does not indicate an endorsement
by the authors or McGraw-Hill Ryerson, and McGraw-Hill Ryerson does
not guarantee the accuracy of the information presented at these sites.

ISBN-13: 978-1-259-65477-0 ISBN-10: 1-259-65477-X

1 2 3 4 5 6 7 8 9 M 23 22 21 20 19

Printed and bound in Canada.

Care has been taken to trace ownership of copyright material contained in
this text; however, the publisher will welcome any information that enables
them to rectify any reference or credit for subsequent editions.

Product Director: *Rhondda McNabb* Portfolio Manager: *Alex Campbell*
Marketing Manager: *Patti Rozakos* Content Developer: *Shalini Khanna*
Portfolio Associate: *Tatiana Sevciuc* Supervising Editor: *Jack Whelan*
Photo/Permissions Research: *Photo Affairs, Inc.* Copy Editor: *Michael Kelly*
Plant Production Coordinator: *Joelle McIntyre* Manufacturing
Production Coordinator: *Jason Stubner* Cover Design: *Liz Harasymczuk*
Cover Image: © ehtesham/Shutterstock Interior Design: *Liz Harasymczuk*
Page Layout: *Aptara®*, ,Inc. Printer: *Marquis*

Dedication

To my wife and family, For their unwavering support and trust. —R.A.M.

To Kathleen D. Vohs, For your mentorship and encouragement. —C.D.R.

For Ingrid and Pierre. —P.C.C.

About the Authors

Paul C. Cozby is emeritus professor of psychology at California State University, Fullerton. Dr. Cozby was an undergraduate at the University of California, Riverside, and received his Ph.D. in psychology from the University of Minnesota. He is a fellow of the American Psychological Association and a member of the Association for Psychological Science; he has served as officer of the Society for Computers in Psychology. He is executive officer of the Western Psychological Association. He is the author of *Using Computers in the Behavioral Sciences* and co-editor with Daniel Perlman of *Social Psychology*.

Raymond A. Mar is a professor of psychology at York University, where he teaches research methods to undergraduates and graduate students. Dr. Mar completed his undergraduate and graduate training at the University of Toronto ultimately receiving his Ph.D. in social/personality psychology. His lab conducts research on how imagined experiences affect real-world cognition and behaviour, with a focus on experiences with fictional stories (e.g., novels, television shows, graphic novels, video games). This work has been published in journals such as the *Annual Review of Psychology* and *Discourse Processes* (for details, visit <http://www.yorku.ca/mar/>). Dr. Mar was awarded the Tom Trabasso Young Investigator Award by the Society for Text and Discourse, and the Friedrich Wilhelm Bessel Research Award by the Alexander von Humboldt Foundation in Germany.

Catherine D. Rawn is a professor of teaching in the department of psychology at the University of British Columbia. Dr. Rawn was an undergraduate psychology major at the University of Waterloo and received her master of arts in social/personality psychology at the University of British Columbia. She continued at UBC and received her Ph.D. in social/personality psychology with a minor in quantitative methods. Her publications appear in journals such as *Teaching of Psychology* and *Personality and Social Psychology Review*. She regularly teaches undergraduates research methods, introductory statistics, and introductory

psychology. She also facilitates professional development in teaching for graduate students and faculty.

Brief Contents

1. Preface xii
2. 1 Scientific Understanding of Behaviour 1
3. 2 Where to Start 15
4. 3 Ethical Research 36
5. 4 Research Design Fundamentals 62
6. 5 Measurement 89
7. 6 Observational Methods 106
8. 7 Survey Research: Asking People about Themselves 124
9. 8 Experimental Design 149
10. 9 Conducting Studies 164
11. 10 Research Designs for Special Circumstances 184
12. 11 Complex Experimental Designs 209
13. 12 Descriptive Statistics: Describing Variables and the Relations among Them 226
14. 13 Inferential Statistics: Making Inferences about Populations Based on Our Samples 250
15. 14 Generalizing Results 276
16. Appendix A: Writing Research Reports in APA Style 294

17. Appendix B: Statistical Tests 334
18. Appendix C: Statistical Tables 353
19. Appendix D: How to Conduct a PsycINFO Search 359
20. Appendix E: Constructing a Latin Square 363
21. Glossary GL-1
22. References RE-1
23. Index IN-1

Contents

Preface xii

1. 1 Scientific Understanding of Behaviour 1
1. Why Study Research Methods? 2
2. Methods of Acquiring Knowledge 3
 1. Intuition 3
 2. Authority 4
 3. The Scientific Method: Be Skeptical, Seek Empirical Data 4
 4. Science as a Way to Ask and Answer Questions 6
3. Goals of Scientific Research in Psychology 8
 1. Describing Behaviour 8
 2. Predicting Behaviour 9
 3. Determining the Causes of Behaviour 9
 4. Explaining Behaviour 10
4. Basic and Applied Research 11
 1. Basic Research 11
 2. Applied Research 11
 3. Integrating Basic and Applied Research 12

- 5. Study Terms* 13
- 6. Review Questions* 13
- 7. Deepen Your Understanding* 14
 - 1. 2 Where to Start 15
 - 1. Where Do Research Ideas Come From? 16
 - 1. Questioning Common Assumptions 16
 - 2. Observation of the World around Us 16
 - 3. Practical Problems 17
 - 4. Theories 18
 - 5. Past Research 19
 - 2. How Do We Find Out What Is Already Known? 20
 - 1. What to Expect in a Research Article 20
 - 2. Other Types of Articles: Literature Reviews and Meta-Analyses 24
 - 3. Reading Articles 25
 - 4. Where Are These Articles Published? An Orientation to Journals and Finding Articles 25
 - 3. Developing Hypotheses and Predictions 30
 - 4. *Study Terms* 34
 - 5. Review Questions* 34
 - 6. Deepen Your Understanding* 34

1.	3 Ethical Research 36
1.	Were Milgram's Obedience Experiments Ethical? 37
2.	Ethical Research in Canada 38
1.	The Tri-Council and Its Policy Statement 38
2.	Historical, Legal, and International Context 38
3.	Core Principles Guiding Research with Human Participants 39
3.	Designing Research to Uphold the Core Principles 39
1.	Promote Concern for Welfare by Minimizing Risks and Maximizing Benefits 40
2.	Promote Respect for Persons through Informed Consent 43
3.	Promote Justice by Involving People Equitably in Research 50
4.	Evaluating the Ethics of Research with Human Participants 50
4.	Monitoring Ethical Standards at Each Institution 51
1.	Exempt Research 52
2.	Minimal Risk Research 52
3.	Greater Than Minimal Risk Research 52
5.	Ethics and Animal Research 53
6.	Professional Ethics in Academic Life 54
1.	Ethics Codes of the APA and CPA 54
2.	Scientific Misconduct and Publication Ethics 55

3. Plagiarism and the Integrity of Academic Communication 57

7. *Study Terms* 60

8. *Review Questions* 60

9. *Deepen Your Understanding* 60

1. 4 Research Design Fundamentals 62

1. Introduction to Basic Research Design 63

1. Variables 63

2. Two Basic Research Designs 63

3. Operationally Defining Variables: Turning Hypotheses into Predictions 65

2. Non-experimental Method 67

1. Relationships between Variables 67

2. Interpreting the Results of Non-experimental Designs 73

3. Experimental Method 76

1. Designing Experiments That Allow for Causal Inferences 77

4. Choosing a Method: Advantages of Multiple Methods 81

1. Artificiality of Experiments 81

2. Ethical and Practical Considerations 83

3. Describing Behaviour 83

4. Predicting Future Behaviour 84

5. Advantages of Multiple Methods 84

- 5. Study Terms* 86
- 6. Review Questions* 86
- 7. Deepen Your Understanding* 87
 - 1. 5 Measurement 89
 - 1. Self-Report Measures 90
 - 2. Reliability 90
 - 1. Test-Retest Reliability 93
 - 2. Internal Consistency Reliability 94
 - 3. Inter-rater Reliability 95
 - 4. Reliability and Accuracy of Measures 95
 - 3. Validity of Measures 95
 - 1. Indicators of Construct Validity 96
 - 4. Reactivity of Measures 99
 - 5. Variables and Measurement Scales 100
 - 1. Nominal Scales 101
 - 2. Ordinal Scales 101
 - 3. Interval Scales 101
 - 4. Ratio Scales 102
 - 5. The Importance of the Measurement Scales 102

Page vii

- 6. Study Terms* 104
- 7. Review Questions* 104
- 8. Deepen Your Understanding* 104
 - 1. 6 Observational Methods 106
 - 1. Quantitative and Qualitative Approaches 107
 - 2. Naturalistic Observation 108
 - 1. Issues in Naturalistic Observation 110
 - 3. Systematic Observation 112
 - 1. Coding Schemes 113
 - 2. Issues in Systematic Observation 114
 - 4. Case Studies 115
 - 5. Archival Research 117
 - 1. Census Data or Statistical Records 117
 - 2. Survey Archives 118
 - 3. Written Records and Mass Media 119
 - 4. Working with Archival Data: Content Analysis and Interpretation 120
 - 6. *Study Terms* 122
 - 7. Review Questions* 122
 - 8. Deepen Your Understanding* 123
 - 1. 7 Survey Research: Asking People about Themselves 124

1. Why Conduct Surveys? 125
 1. Response Bias in Survey Research 127
2. Constructing Good Questions 128
 1. Defining the Research Objectives 128
 2. Question Wording 129
3. Responses to Questions: What Kind of Data Are You Seeking? 132
 1. Closed- versus Open-Ended Questions 132
 2. Rating Scales for Closed-Ended Questions 132
4. Finalizing the Questionnaire 135
 1. Formatting the Questionnaire 135
 2. Refining Questions 136
5. Administering Surveys 136
 1. Questionnaires 136
 2. Interviews 138
6. Interpreting Survey Results: Consider the Sample 139
 1. Population and Samples 140
 2. For More Precise Estimates, Use a Larger Sample 141
 3. To Describe a Specific Population, Sample Thoroughly 141
7. Sampling Techniques 143

Page viii

1. Probability Sampling 143
 2. Non-probability Sampling 145
 3. Reasons for Using Convenience Samples 146
 8. *Study Terms* 147
 9. *Review Questions* 147
10. *Deepen Your Understanding* 148
1. 8 Experimental Design 149
 1. Confounding and Internal Validity 150
 2. Planning a Basic Experiment 151
 3. Between-Subjects Experiments 153
 1. Pretest-Posttest Design 154
 2. Matched Pairs Design 155
 4. Within-Subjects Experiments 156
 1. Advantages and Disadvantages of the Within-Subjects Design 156
 2. Counterbalancing 158
 3. Time Interval between Treatments 160
 4. Choosing between Between-Subjects and Within-Subjects Designs 160
 5. *Study Terms* 162
 6. *Review Questions* 162

- 7. Deepen Your Understanding* 162
1. 9 Conducting Studies 164
 1. Finalizing a Study Design 165
 1. Options for Manipulating the Independent Variable in Experiments 165
 2. Additional Considerations When Manipulating the Independent Variable 167
 3. Options for Measuring Variables 169
 4. Additional Considerations When Measuring Variables 172
 5. Setting the Stage 173
 2. Advanced Considerations for Ensuring Control 174
 1. Controlling for Participant Expectations 174
 2. Controlling for Experimenter Expectations 175
 3. Seeking Ethics Approval 177
 1. Selecting Research Participants 177
 2. Planning the Debriefing 178
 4. Collecting Data 179
 1. Pilot Studies 180
 2. Researcher Commitments 180
 5. What Comes Next? 180
 1. Analyzing and Interpreting Results 180

- 2. Communicating Research to Others 180
- 6. *Study Terms* 182
- 7. *Review Questions* 182
- 8. *Deepen Your Understanding* 183
 - 1. 10 Research Designs for Special Circumstances 184
 - 1. Program Evaluation 185
 - 2. Quasi-Experimental Designs 187
 - 1. One-Group Posttest-Only Design 189
 - 2. One-Group Pretest-Posttest Design 190
 - 3. Threats to Internal Validity 190
 - 4. Non-equivalent Control Group Design 195
 - 5. Non-equivalent Control Group Pretest-Posttest Design 195
 - 6. Interrupted Time Series Design 196
 - 7. Control Series Design 197
 - 8. Summing Up Quasi-Experimental Designs 198
 - 3. Single Case Experimental Designs 199
 - 1. Reversal Designs 199
 - 2. Multiple Baseline Designs 200
 - 3. Replications in Single Case Designs 202

Page ix

4. Developmental Research Designs 202
 1. Longitudinal Method 203
 2. Cross-Sectional Method 204
 3. Comparing Longitudinal and Cross-Sectional Methods 204
 4. Sequential Method 205
5. *Study Terms* 206
6. *Review Questions* 207
7. *Deepen Your Understanding* 207
 1. 11 Complex Experimental Designs 209
 1. An Independent Variable with More Than Two Levels 210
 2. An Experiment with More Than One Independent Variable: Factorial Designs 212
 1. Interpreting Factorial Designs 213
 2. Interactions Illuminate Moderator Variables 215
 3. Depicting Possible Outcomes of a 2×2 Factorial Design Using Tables and Graphs 215
 4. Breaking Down Interactions into Simple Main Effects 218
 3. Variations on 2×2 Factorial Designs 218
 1. Factorial Designs with Manipulated and Non-manipulated Variables 218
 2. Assignment Procedures and Sample Size 219

4. Increasing the Complexity of Factorial Designs 221
 1. Beyond Two Levels per Independent Variable 221
 2. Beyond Two Independent Variables 222
5. *Study Terms* 224
6. *Review Questions* 224
7. *Deepen Your Understanding* 225
 1. 12 Descriptive Statistics: Describing Variables and the Relations among Them 226
 1. Revisiting Scales of Measurement 227
 2. Describing Each Variable 228
 1. Graphing Frequency Distributions 228
 2. Descriptive Statistics 231
 3. Describing Relationships Involving Nominal Variables 234
 1. Comparing Groups of Participants 234
 2. Graphing Nominal Data 235
 3. Describing Effect-Size between Two Groups 235
 4. Describing Relationships among Continuous Variables: Correlating Two Variables 237
 1. Interpreting the Pearson r Correlation Coefficient 237
 2. Scatterplots 238

Page x

- 3. Important Considerations 240
- 4. Correlation Coefficients as Effect-Sizes 241
- 5. Describing Relationships among Continuous Variables: Increasing Complexity 242
 - 1. The Regression Equation 242
 - 2. Multiple Correlation and Multiple Regression 243
 - 3. Integrating Results from Different Analyses 245
 - 4. Partial Correlation and the Third-Variable Problem 245
 - 5. Advanced Modelling Techniques 246
- 6. Combining Descriptive and Inferential Statistics 247
- 7. *Study Terms* 247
- 8. *Review Questions* 248
- 9. *Deepen Your Understanding* 248
 - 1. 13 Inferential Statistics: Making Inferences about Populations Based on Our Samples 250
 - 1. Inferential Statistics: Using Samples to Make Inferences about Populations 251
 - 1. Inferential Statistics: Ruling Out Chance 251
 - 2. Statistical Significance: An Overview 252
 - 2. Null and Research Hypotheses 253
 - 3. Probability and Sampling Distributions 253

1. Probability: The Case of Mind Reading 254
2. Sampling Distributions 254
3. Sample Size 256
4. How “Unlikely” Is Enough? Choosing a Statistical Significance Level (Alpha) 256
4. Example Statistical Tests 257
 1. The *t*-Test: Comparing Two Means 257
 2. The *F* Test: Used When Comparing Three or More Group Means 259
 3. Statistical Significance of a Pearson *r* Correlation Coefficient 259
5. We Made a Decision about the Null Hypothesis, but We Might Be Wrong! Investigating Type I and Type II Errors 260
 1. Correct Decisions 260
 2. Type I Errors 260
 3. Type II Errors 261
 4. The Everyday Context of Type I and Type II Errors 261
 5. Type I and Type II Errors in the Published Research Literature 263
6. Interpreting Statistically Non-significant Results 264
7. Choosing a Sample Size: Power Analysis 265
8. Analyzing Data Using Statistics Software 267
9. Selecting the Appropriate Statistical Test 269

1. Research Studying Two Variables 269
2. Research with Multiple Independent or Predictor Variables 269

Page xi

10. Integrating Descriptive and Inferential Statistics 270
 1. Effect-Size 271
 2. Confidence Intervals and Statistical Significance 272
 3. Conclusion Validity 273
11. The Importance of Replications 273
12. *Study Terms* 273
13. *Review Questions* 274
14. *Deepen Your Understanding* 274
 1. Generalizing Results 276
 1. Challenges to Generalizing Results 277
 1. Can Results Generalize to Other Populations? 277
 2. Can Results Generalize Beyond the Specific Study Situation? 280
 2. Solutions to Generalizing Results 281
 1. Replicate the Study 281
 2. Consider Different Populations 285
 3. Rely on Multiple Studies to Draw Conclusions: Literature Reviews and Meta-Analyses 288
 3. Generalizing Your Knowledge beyond This Book 289

1. Recognize and Use Your New Knowledge 290
 2. Stay Connected to Building a Better Psychological Science 290
 3. Use Research to Improve Lives 291
 4. *Study Terms* 292
 5. *Review Questions* 292
 6. *Deepen Your Understanding* 293
- Appendix A: Writing Research Reports in APA Style 294
- Appendix B: Statistical Tests 334
- Appendix C: Statistical Tables 353
- Appendix D: How to Conduct a PsycINFO Search 359
- Appendix E: Constructing a Latin Square 363
- Glossary GL-1
- References RE-1
- Index IN-1

Preface

Welcome to the third Canadian edition of *Methods in Behavioural Research*! When I first began teaching research methods, I found it immediately rewarding for two main reasons. The first reason is that it expanded my horizons, adding nuance and breadth to my conception of science, all of which I hope to pass on to the readers. The second reason is that I know I am teaching a valuable skill that will take my students far in life, no matter what direction they choose. Scientific thinking is not a strict set of routine procedures that are useful only for those who become scientists. It is a way of thinking about the world that acknowledges its messy complexity and helps us reason about the best way to solve problems and evaluate evidence, all in the service of making informed decisions. Just like in many aspects of life, there is not one ultimate way of doing things; there are many options, each with strengths and weaknesses in different contexts, and our goal is to try to identify the best option given the circumstances.

Features of This Canadian Edition

In this latest edition, I have focused on bolstering the Canadian content, increasing the opportunity for interactive engagement with the material, updating the treatment of statistical issues, and improving the clarity and concision of communication throughout. Our text hopes to inspire the next generation of Canadian researchers by highlighting the outstanding research being done in this country. This research is being conducted by individuals who were once undergraduates in a very similar position: embarking on their journey as a scientist. In addition to incorporating Canadian research to illustrate concepts and using culturally relevant examples, I have added a new feature: Student Spotlights. These spotlights introduce research conducted by Canadian undergraduate students, research that has gone on to be published in a peer-reviewed academic journal.

Hopefully these examples not only help you to learn the content of this book but also convince you that you too can become a successful scientist, contributing to our shared understanding of the world.

In addition to including exemplars of successful Canadian scientists (both faculty and students), I have increased the interactivity of this text in order to help you engage with the content directly. Throughout, readers are encouraged to actively engage with ideas via Think About It! boxes as well as participate in Test Yourself! exercises to evaluate current understanding. At the end of many chapters, Illustrative Articles are now featured, in which students learn to acquire and read relevant articles, then answer questions that help to illustrate the main concepts of the chapter.

All of these changes build upon the very solid foundation established in the many previous editions of this text (both Canadian and American). What has made this text a favourite of instructors and students alike is its clarity of expression, its concision, and the reinforcement of key constructs. To the best of my abilities, I have tried to preserve and improve on these positive qualities throughout. For example, glossary definitions in the print edition now appear in the margins, helping students to access the definitions of major ideas quickly and easily. Revisions to the writing have been made throughout with a focus on brevity and clarity. In addition, language has been updated to be more inclusive (e.g., removing mention of gender as a binary construct). The strengths of previous editions—in the form of structured and explicit learning objectives and end-of-chapter review questions—remain.

Page xiii

With respect to content, I have tried to reinforce the idea that the numerous approaches to research are not better or worse than one another: merely different. Each has unique strengths and weaknesses, with different research questions and contexts calling for separate approaches.

Organization & Other changes

The largest organizational change to this edition is that what was formerly Chapter 11, *Research Designs for Special Circumstances*, has now been moved to Chapter 10 (and the former Chapter 10, *Complex Experimental*

Designs, is now moved to Chapter 11). The rationale for this decision is that quasi-experimental designs serve as great illustrations of potential weaknesses in studies that purport to be true experiments. The fundamental issue of validity for an experiment is necessary to grasp before moving on to more complex experimental designs. Importantly, we have now made it clear that these quasi-experiments are not “bad” experiments; they are simply different designs that are often necessary when a true experiment is not possible. However, when a true experiment is possible (and, often, is claimed), one can often identify threats to the validity of this experiment by identifying qualities of quasi-experiments. Otherwise, the organization has largely been kept intact in this edition. What follows is a brief overview of the content for each chapter, along with the changes in this edition.

- **Chapter 1.** *The goals and methods of the scientific method, basic versus applied research.*
 - Added 5 Student Spotlights, 2 Think About It! boxes, and 1 Research Spotlight. Removed dated examples. Updated Figure 1.1.
- **Chapter 2.** *Generating research ideas, finding and reading past research, developing hypotheses.*
 - Added 3 Student Spotlights and an Illustrative Article. Extended discussion of how to approach abstracts. Removed dated examples. Emphasized importance of reading articles critically. Updated table of major journals. Expanded discussion of cited reference searches.
- **Chapter 3.** *Research ethics in Canada, for human and animal research; scientific misconduct.*
 - Added 2 Student Spotlights, 5 Think About It! boxes, and an Illustrative Article. Removed dated examples and discussion.
- **Chapter 4.** *Introducing experimental and non-experimental methods, operationalization, choosing a method, and using multiple methods.*

- Added 3 Student Spotlights, 3 Think About It! boxes, 2 Try it Out! boxes, 3 Test Yourself! boxes, and an Illustrative Article. Expanded on the strengths and weakness of both experimental and non-experimental designs. Clarified and expanded the discussion of lab-field correspondence.
- **Chapter 5.** *Self-report measures, reliability and validity, participant reactivity, different measurement scales.*
 - Added 1 Student Spotlight, 3 Think About It! boxes, and an Illustrative Article. Expanded and clarified the concept of measurement error and its relationship to true scores and reliability. Elaborated on participant reactivity.
- **Chapter 6.** *Quantitative and qualitative approaches to observational methods, case studies and archival research.*
 - Added 3 Student Spotlights, 1 Think About It! box, 1 Try it Out! box, 1 Test Yourself! box, and an Illustrative Article. Clarified how qualitative and quantitative approaches map onto the goals of science and different research questions. Updated discussion of technology, including tools for analyzing different forms of archival data. Page xiv
- **Chapter 7.** *Correlational survey research, constructing items and response scales, sampling.*
 - Added 2 Student Spotlights, 1 Think About It! box, and 1 Try it Out! box. Updated and expanded discussion of inaccurate responding and socially desirable responding, including mention of tools for catching inattentive responders. Introduced the term acquiescence bias, and discussion of how samples may or may not generalize to populations. Expanded on participant fatigue. Included the practical definition of a confidence interval, distinguishing this from the theoretical or statistical definition, and removing prior incorrect definition. Defined science as a process of accumulating knowledge, with no single study being authoritative.

- **Chapter 8.** *Between-subjects and within-subjects experiments, internal validity and confounds.*
 - Added 1 Student Spotlight, 1 Think About It! box, and an Illustrative Article. Changed key terminology to between-subjects and within-subjects designs (from independent-groups and repeated-measures designs). The matched pairs design is now discussed within the between-subjects designs section. Updated examples and discussion of sample size recommendations for random assignment to conditions. Changed the term mortality to selective attrition. Removed discussion of a Solomon four-group design.
- **Chapter 9.** *Manipulating and measuring variables, controls for participant and experimenter variables, conducting studies.*
 - Added 2 Student Spotlights, 1 Think About It! box, and an Illustrative Article. Clarified the difference between experimental realism and mundane realism. Added mention of the risks of manipulation checks. Clarified discussion of fMRI. Elaborated on diagnosing ceiling and floor effects. Added discussion of selective reporting as an unethical practice. Clarified that placebo effects can occur without drug administration. Elaborated on the importance of debriefings.
- **Chapter 10.** *Quasi-experiments, case-studies, developmental designs, program evaluation.*
 - Formerly Chapter 11. Added 4 Think About It! boxes and an Illustrative Article. Moved discussion of quasi-experiments closer to the beginning of this chapter. Clarified that quasi-experiments are *not* true experiments that lack internal validity, but rather unique designs that are sometimes necessary. Added an exercise to consider situations in which quasi-experiments are necessary. Used quasi-experiments to illustrate flaws in designs that claim to be true experiments. Clarified why regression to the mean is related to reliability and measurement error.

- **Chapter 11.** *Multi-level independent variables, factorial experimental designs, IV × PV designs.*
 - Formerly Chapter 10. Added 2 Try it Out! boxes, 2 Think About It! boxes, and an Illustrative Article. Clarified the example for the utility of additional control conditions and the meaning of statistical significance.
- **Chapter 12.** *Descriptive statistics, effect-sizes, graphing data, describing relationships between variables.*
 - Added 2 Student Spotlights, 1 Think About It! box, 3 Try it Out! boxes, and 2 Test Yourself! Boxes. Added explanation of the normal distribution and its importance for parametric statistics. Added formula for calculating Cohen's d and practice exercises. Expanded on when you use each measure of central tendency and why. Updated guidelines for interpreting effect-size magnitude. Clarified the relationship between effect-size, practical significance, and statistical significance. Removed discussion of how regressions are framed as predictions whereas correlations are not, since both can be used for prediction. Expanded on common misinterpretations of Pearson r .Page xv
- **Chapter 13.** *Inferential statistics, statistical significance, Type I and Type II errors, alternatives to null-hypothesis significance testing (NHST).*
 - Added 2 Think About It! boxes and 3 Try it Out! boxes. Clarified that NHST is just one option for inferential statistics, clarified common misconceptions of NHST, and expanded discussion of the controversy surrounding this approach. Removed manual calculation of the t -value in favour of emphasizing conceptual understanding. Removed discussion of one-tailed tests, as this procedure is now known to be inappropriate. Greatly truncated the discussion of looking up critical values, as this is an outdated practice, replacing this with discussion of p -values and comparison to alpha instead. Introduced Bayesian statistics as an appropriate way to evaluate evidence in favour of the null. Added

references to introductory articles on Bayesian statistics, and introduced free software for performing Bayesian analyses. Included the fact that the parametric statistics discussed rely on an assumption of sampling from normally distributed populations. Added list of resources for learning the statistical software R.

- **Chapter 14.** *Challenges of generalizing results, importance of replication, literature reviews and meta-analyses, applying your knowledge.*
 - Added 2 Student Spotlights and an Illustrative Article. Expanded and clarified how samples relate to the populations from which they were drawn, and generalizing to other populations. Expanded discussion of whether laboratory studies and field experiments converge on similar results. Clarified when replication failures can be informative and how conceptual replications can be problematic.

Award-Winning Technology



McGraw-Hill Connect® is an award-winning digital teaching and learning solution that empowers students to achieve better outcomes and enables instructors to improve efficiency with course management.

Within Connect, students have access to SmartBook®, McGraw-Hill's adaptive learning and reading resource. SmartBook prompts students with questions based on the material they are studying. By assessing individual answers, SmartBook learns what each student knows and identifies which topics they need to practise, giving each student a personalized learning experience and path to success.

Connect's key features also include analytics and reporting, simple assignment management, smart grading, the opportunity to post your own resources, and the Connect Instructor Library, a repository for additional resources to improve student engagement in and out of the classroom.

Instructor Resources for *Methods in Behavioural Research*, Third Canadian Edition

- **Instructors Manual.** This in-depth Instructors Manual offers numerous student activities and assignment suggestions as well as demonstrations, discussions topics, reference articles, and sample answers for questions in the text.
- **Test Bank.** Within Connect, instructors can easily create automatically graded assessments from a comprehensive test bank featuring multiple question types and randomized question order.Page xvi
- **Connect Assignments.** The assignable resources, such as NewsFlash articles and Test questions, can be used to create assignments.
- **Power of Process.** New to the third edition, Power of Process for *Methods in Behavioural Research* helps students improve critical-thinking skills and allows instructors to assess these skills efficiently and effectively in an online environment. Available through Connect, preloaded journal articles are available for instructors to assign. Using a scaffolded framework such as understanding, synthesizing, and analyzing, Power of Process moves students toward higher-level thinking and analysis.
- **Microsoft® PowerPoint® Presentations.** The customizable PowerPoint presentations represent the key concepts in each chapter.

Acknowledgments

There are a great many people who deserve thanks for helping to produce this third Canadian edition. I am, first and foremost, deeply indebted to the previous authors: Paul C. Cozby, the original author of the American

edition; all others who worked on those editions; and especially Catherine D. Rawn, who did an absolutely fantastic job of preparing the previous two Canadian editions. I would also like to thank all at McGraw-Hill Education who worked on this book, especially portfolio manager, Alex Campbell, content developer, Shalini Khanna, and supervising editor, Jack Whelan. Michael Kelly (Good Eye Editorial Services) provided invaluable copy editing, and Steve Rouben (Photo Affairs) excelled at the necessary photo research. My undergraduate research assistant, Alma Rahimi, deserves special thanks for helping to compile the database of student publications that was used to prepare the Student Spotlights.

The following people are also thanked for generously providing reviews for this book:

1. Craig Blatz *Grant MacEwan University*
2. Connie Boudens *University of Toronto, Scarborough*
3. Patrick Brown *University of Western Ontario*
4. Keith Busby *University of Ottawa*
5. Laura Dane *Douglas College*
6. Lucie Kocum *Saint Mary's University*
7. Guy Lacroix *Carleton University*
8. Chris Montoya *Thompson Rivers University*
9. Margarete Wolfram *York University*

Nothing I do would be possible without my wife, for whom I am ever grateful. My family and friends have always supported me and are also deserving of much thanks. Over my career, I have learned so much from my students, both undergraduate and graduate students, as well as many of my colleagues, and much of this knowledge has found its way into this book. Lastly, I would like to thank you, the reader, for reading this book and considering its lessons, and for bothering to read this lengthy preface!

—Raymond A. Mar



Effective. Efficient. Easy to Use.

McGraw-Hill Connect is an award-winning digital teaching and learning solution that empowers students to achieve better outcomes and enables instructors to improve course-management efficiency.



Personalized & Adaptive Learning

Connect's integrated SmartBook helps students study more efficiently, highlighting where in the text to focus and asking review questions to give each student a personalized learning experience and path to success.



High-Quality Course Material

Our trusted solutions are designed to help students actively engage in course content and develop critical higher-level thinking skills, while offering you the flexibility to tailor your course to meet your needs.



Analytics & Reporting

Monitor progress and improve focus with Connect's visual and actionable dashboards. Reporting features empower instructors and students with real-time performance analytics.



Seamless Integration

Link your Learning Management System with Connect for single sign-on and gradebook synchronization, with all-in-one ease for you and your students.

Impact of Connect on Pass Rates



Without Connect



With Connect

SMARTBOOK

NEW SmartBook 2.0 builds on our market-leading adaptive technology with enhanced capabilities and a streamlined interface that deliver a more usable, accessible and mobile learning experience for both students and instructors.



Available on mobile smart devices – with both online and offline access – the ReadAnywhere app lets students study anywhere, anytime.

SUPPORT AT EVERY STEP

McGraw-Hill ensures you are supported every step of the way. From course design and set up, to instructor training, LMS integration and ongoing support, your Digital Success Consultant is there to make your course as effective as possible.

Learn more about Connect at mheducation.ca

Scientific Understanding of Behaviour



©Mark Bridger/Shutterstock

We think of owls as wise and intelligent, but in fact they're not particularly quick learners. This is why we need science: to question our assumptions and pursue truth through systematic observation.

Learning Objectives

Keep these learning objectives in mind as you read to help you identify the most critical information in this chapter.

By the end of this chapter, you should be able to:

1. LO1 Explain reasons why understanding research methods is important.
2. LO2 Describe the scientific approach to learning about behaviour and contrast it with pseudoscience.
3. LO3 Define and give examples of the four goals of scientific research in psychology.
4. LO4 Compare and contrast basic and applied research.

Page 2What makes people happy? How do we remember things, what causes us to forget, and how can memory be improved? What are the effects of stress on physical health and relationships? How do early childhood experiences affect later development? What are the best ways to treat depression? How can we reduce prejudice and conflict? Curiosity about questions like these is probably the most important reason many students decide to take courses in the behavioural sciences. Scientific research provides us with a way to gather evidence that can shape our beliefs about the answers to such questions. Throughout this book, we will examine the methods employed for scientific research in the behavioural sciences. In this introductory chapter, we will focus on the ways in which knowledge of research methods can be useful for understanding the world around us. Further, we will review the characteristics of a scientific approach to the study of behaviour and the general types of research questions that concern behavioural scientists.



LO1 Why Study Research Methods?

Understanding research methods can help you become an informed consumer of news, health care, products, and services. Scientific research is frequently reported by news organizations, popular magazines, bloggers, and advertisers. Headlines for these stories may make bold claims and ask provocative questions, such as “Study finds that lonely people use Facebook all the time,” “Will getting a dog help you live longer?” and “When drugs and therapy don’t cure depression, running will.” In addition, we often hear about survey results that draw conclusions about a group’s beliefs and attitudes. How do you evaluate such reports? Do you simply accept the findings because they seem scientific? Can you detect pseudoscientific claims (as we will explore [later](#) in this chapter)? A background in research methods will help you to read these reports critically, evaluate the methods, and decide whether the conclusions and assertions being made are appropriate and justifiable.

Understanding research methods can give you a competitive edge for various careers. Many occupations require the ability to interpret, appropriately apply, and conduct solid research. For example, mental health professionals must make decisions about treatment methods, medications,

and testing procedures; this requires the ability to read the relevant research literature and apply it effectively. Similarly, people who work in business frequently rely on research to make decisions about marketing strategies, ways of improving employee productivity or morale, and methods of selecting and training new employees. Educators must also keep up with research, for topics such as the effectiveness of different teaching strategies or programs for students with special challenges. Others are engaged with program evaluation, conducting research to evaluate the efficacy of government and other programs to ensure that funding is well-spent (elaborated on [later](#) in this chapter). Knowledge of research methods and the ability to evaluate research reports are essential for these and other careers.

Page 3

Understanding research methods can help you be an informed and engaged citizen and participate in debates regarding public policy. Legislators and political leaders at all levels of government often take political positions and propose legislation based on research findings. Research can also influence legal practices and decisions. For example, numerous wrongful convictions triggered the use of psychological research to inform police investigation and courtroom procedures (Public Prosecution Service of Canada [PPSC], 2011; U.S. Department of Justice, 1999; Wells, 2001; Yarmey, 2003). In one case, Thomas Sophonow was wrongfully convicted of murder by a Manitoba jury in 1983. After serving four years of his life sentence, his conviction was overturned. In the inquiry that followed (Cory, 2001), a retired Supreme Court of Canada judge used psychological science as the basis for numerous recommendations to prevent future wrongful convictions. One of the studies that influenced these recommendations was conducted at Queen's University and showed that people make fewer false identifications of suspects when they are presented with a set of photographs one at a time, rather than simultaneously (Lindsay & Wells, 1985; see also Steblay, Dysart, Fulero, & Lindsay, 2001). Since this inquiry, police have been required to present photographs sequentially when asking eyewitnesses to identify a suspect (PPSC, 2011). Another way that psychologists influence judicial decisions is by providing expert testimony and consultation on a variety of issues, including domestic violence (e.g., *R. v. Lavallee*, 1990), risk for violence (e.g., *R. v. Berikoff*, 2000), and memories retrieved through hypnosis (e.g., *R. v. Trochym*, 2007).

Understanding research methods can help you evaluate programs in your community that you might want to participate in or even implement. There exist many different programs to provide assistance to different groups. For example, there are programs to enhance parenting skills for parents of aggressive and antisocial youth (Moretti & Obsuth, 2009), to help reduce behaviours that raise your risk of contracting HIV, and to teach employees and students how to reduce the effects of stress. We need to be able to determine whether these programs are successfully meeting their goals, and the application of research methods helps us do just that.

Methods of Acquiring Knowledge

We opened this chapter with several questions about human behaviour and suggested that scientific research is a valuable means of gathering information about the answers to these questions. How does the scientific approach differ from other ways of learning about behaviour? People have always observed the world around them and sought explanations for what they see and experience. In this quest, humans often rely solely on intuition and authority, but these can lead to biased conclusions. Science offers a way to try to avoid some of these biases by systematically seeking high-quality evidence.

Intuition

Many of us have heard about someone who, after years of actively looking for a long-term romantic partner, stops looking for love. Then, soon after, this same person happens to find the love of their life! Anecdotes like this contribute to a common belief that love arrives when one is not looking for it. This seems intuitively reasonable, and people can easily create an explanation for why this is the case (see Gilovich, 1991). Perhaps stopping the hunt reduces a major source of stress, this reduction in stress increases our confidence in social interactions, which in turn makes us more desirable to potential partners.

This example illustrates the use of *intuition* based on anecdotal evidence to draw general conclusions. When you rely on intuition, you accept unquestioningly what your personal judgment or a single story about one person's experience tells you about the world. The intuitive approach takes many forms. Often, it involves finding an explanation for our own or others' behaviours. For example, you might develop an explanation for why you keep having conflict with a fellow student, such as "that other person is jealous of my intelligence." Other times, intuition is used to explain intriguing events that you simply observe in the world, as in the case of love arriving when you stop looking for it.

One problem with intuition is that many cognitive and motivational biases affect our perceptions, which means we can arrive at mistaken conclusions (cf., Gilovich, 1991; Nisbett & Ross, 1980; Nisbett & Wilson, 1977). So why do we believe that no longer looking for love leads to finding it? Most likely it is because of a cognitive bias called *illusory correlation*: When two events occur

closely in time, this draws our attention, and we often conclude that one must cause the other. In this example, when a decision to stop looking for love is followed closely by finding a long-term mate, our attention is drawn to the situation, and we see them as being causally related. This is true even when it might just be a coincidence. But when a decision to stop looking is *not* closely followed by finding a long-term mate, we don't notice this non-event. Therefore, we are biased to conclude that there must be a causal connection between these things, when in fact no such relationship exists. Illusory correlations are also likely to occur when we are highly motivated to believe that a certain causal relationship is true. If we already believe that not looking for love is the key to finding it, these examples are going to jump out at us even more. Although this way of thinking comes naturally to us as humans, it can lead us to make inaccurate conclusions. A scientific approach tries to overcome this biased way of thinking, and requires much more rigorous evidence before drawing conclusions.

Authority

Other sources of knowledge about the world are various forms of authority. When we make a decision based on *authority*, we place our trust in someone else who we think knows more than we do. When we were young, we likely trusted our parents to know what we should do and what was true about the world. As adults, people tend to trust other authorities such as doctors, especially if they view that doctor as a specialist in the area (Barnoy, Ofra, & Bar-Tal, 2012). Such blind trust in medical authority can be problematic because many health care workers (and patients alike) are prone to drawing incorrect conclusions from statistics regarding health (Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, & Woloshin, 2007). Similarly, many people readily accept anything they encounter from the news media, books, government officials, or religious figures. They believe that the statements of such authorities must be true. Advertisers know this and therefore use authority figures to sell products. The problem, of course, is that the statements by any particular authority may not be true. The scientific approach rejects the notion that one can accept on faith the statements of any authority. The scientific approach is to require lots of evidence, and good quality evidence, before coming to any conclusion.



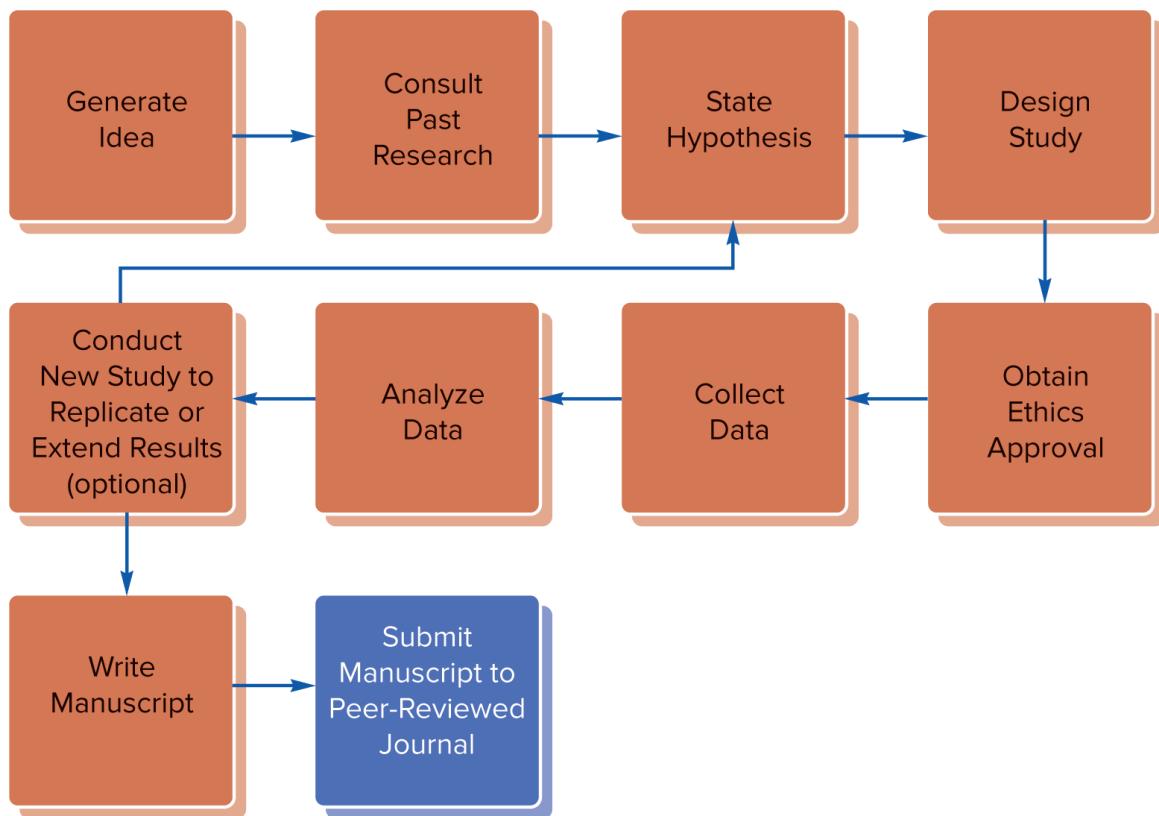
LO2 The Scientific Method: Be Skeptical, Seek Empirical Data

The scientific method of acquiring knowledge acknowledges that both intuition and authority can be useful sources of initial ideas about behaviour. However, the scientific approach does not accept these ideas as truth without further evidence. Being a scientist means not accepting anyone else's intuitions or conclusions without first evaluating the evidence. And this includes our own intuitions and ideas as well. Ideas must be evaluated on the basis of results from structured investigations. Throughout this book, we invite you to try out a mindset of [scientific skepticism](#) (if you haven't already!). Recognize that our own ideas are just as likely to be as wrong as anyone else's, and question any pronouncements of truth, regardless of the prestige or authority associated with the source.

If scientific skepticism involves rejecting intuition and the blind acceptance of authority as ways of knowing about the world, how do we gain knowledge about the world? The fundamental characteristic of the scientific method is [empiricism](#): gaining knowledge based on structured, systematic observations of the world. Although it sounds simple, the process of making systematic observations (i.e., conducting research) is complex and involves many steps. In its basic form, a scientist develops a hypothesis (an idea that might be true; see [Chapter 2](#)), then carefully collects data relevant to this hypothesis, and then evaluates whether the data is consistent or inconsistent with this hypothesis. If the data match the hypothesis, then we have acquired some evidence that the hypothesis might accurately reflect the nature of the world. See [Figure 1.1](#) for an overview of the many steps involved in conducting research. Although all these steps may seem

daunting, this book will help you learn how to develop theories and hypotheses, design studies, collect and evaluate data, and write up the results for publication.

Figure 1.1 Overview of the process of conducting research, scientist's perspective



Think about It!

What step in this process sounds like the most fun for you?

Thousands of individual scientists worldwide—in disciplines as varied as psychology and physics—use the scientific method to understand the world. Regardless of the topic being studied and the specific procedures used, there are some broad characteristics that guide the ideal process of scientific inquiry (Goodstein, 2011; Merton, 1973). It is important to note at the outset, however, that scientists are still human and science is an imperfect enterprise. As a result,

these ideals are not always met in reality. They can be perverted by other pressures in society, such as the pressure to publish many studies to build one's career, as well as pressures from within, like egotism and a desire to be viewed as important by others. The risk of science itself falling short of its ideals is one reason why it's important to adopt a skeptical scientific mindset, even towards scientists and their published works. That said, the majority of scientists across various disciplines agree that the following four norms should characterize scientific inquiry at its best (Anderson, Ronning, DeVries, & Martinson, 2010; Merton, 1973).

1. *Universalism*: Scientific observations are systematic and structured, and evaluated objectively using the accepted methods of the discipline. By relying on empiricism, one group of scientists can publish research and reach one conclusion, another group can disagree and publish their own research that reaches a different conclusion, and the research reported from both sides can be objectively evaluated by others.
Page 6
2. *Communality*: Methods and results are to be shared openly. One major benefit to open reporting is that others can *replicate* a study to see if they get the same results (see [Chapter 14](#); Open Science Collaboration, 2013). Replications help to ensure that the effects being reported are not the result of chance, false positives, scientific fraud, or some other reason (see [Chapter 13](#)). Another major benefit to open reporting is that the results of many studies can be combined in meta-analyses, which combine results from many past studies to examine the overall effect (see [Chapter 14](#)). No single study provides a perfect or complete answer to a research question. Meta-analyses, however, attempt to give a bigger picture of what has been discovered across many studies and thus form an important tool for science. This important tool relies crucially on communality (Braver, Thoemmes, & Rosenthal, 2014; Cumming, 2014). To support the ideal of communality, many researchers have begun posting their data and materials online so that others can properly evaluate and/or replicate their research (Nosek et al., 2015).
3. *Disinterestedness*: Ideally, scientists should be making observations that will help them discover accurate things about the world. They collect systematic observations, develop theories to explain these observations, conduct further research to evaluate these theories, and revise their theories as needed when new data demands it. Scientists should be motivated by an honest and

careful quest for truth, and ideally are not motivated by fame, ego, or personal gain.

4. *Organized skepticism*: All new evidence and theories should be evaluated based on scientific merit, even those that challenge one's own work or prior beliefs. Science exists in a free market of ideas, in which research is expected to be critiqued and evaluated. This means that scientists should ideally be critical of work even if it supports their ideas, and also be open to good research contradicting their own ideas. Of all the ideals, organized skepticism is most directly associated with the practice of [*peer review*](#). Before a study is published in a scientific journal, it must be reviewed by other scientists who have the expertise to carefully evaluate the research and recommend whether the research should be published. Although this peer-review process is intended to form a sort of quality control for research, it is far from perfect, and we must still take a skeptical and critical approach to published peer-reviewed work. But doing so requires understanding the methods used in research, which is why it is important to study research methods. Remember that we don't accept conclusions based on authority alone, so we need tools to be able to evaluate science ourselves. This is where knowledge of research methods comes in.

Science as a Way to Ask and Answer Questions

The main advantage of the scientific approach over other ways of knowing about the world is that it strives to provide an objective way to gather, evaluate, and report evidence. In its ideal form, science is an open system that allows ideas to be refuted or supported on the basis of available evidence. Researchers are only interested in [*falsifiable*](#) ideas, those that can be shown to be false or are capable of being refuted (Popper, 1968). Sometimes, scientific evidence is not obtainable—for example, when religions ask us to accept certain beliefs on faith. It is possible to design a study to test whether a belief in god(s) is associated with altruism, but it is not possible to design a study to test whether a god (or gods) actually exists. The former idea is falsifiable, but the latter is not. This doesn't make the question of whether a god (or gods) exists an uninteresting one (far from it!), it just means this is not a falsifiable [*empirical question*](#). Science can only tackle empirical questions. And falsifiable empirical questions are useful even when they are proven false, because this result will spur the development of new and better ideas.

Our emphasis on science is not meant to imply that intuition and authority are unimportant. In fact, scientists often rely upon both for generating research ideas. There is nothing wrong with accepting the assertions of authority, as long as we do not accept them uncritically as scientific evidence. Likewise, there is nothing wrong with having opinions or beliefs as long as they are presented as such: simply opinions or beliefs. When we take on the scientific mindset, we should always ask whether the assertion or opinion can be tested empirically, or whether scientific evidence exists that relates to the opinion.

☆ Student Spotlight: Examining Opinions ☆

It's possible to have opinions on whether different people have different motivations for drinking alcohol, or even thoughts on what personality traits are tied to these different motivations. However, these are just opinions until we study them empirically. Jennifer Theakston, working with Dr. Sherry Stewart at Dalhousie University, researched this question by surveying almost 600 young adult drinkers about their motivations for drinking and their personality traits. What they found, consistent with past work, was that people drink for both external reasons (e.g., conformity to norms, to be social) and internal reasons (e.g., to cope with stress). Moreover, personality traits successfully predicted these different motives. For example, people who were low in trait emotional stability and low in extraversion were more likely to drink in order to cope with stressors. The results of this research now appear in the journal *Personality and Individual Differences* (Theakston, Stewart, Dawson, Knowlden, & Lehman, 2004).

Truly embracing scientific skepticism means questioning the claims of scientists themselves. Scientists can become authorities when they express their ideas because they have some expertise. But this does not mean that their claims should be believed unthinkingly. It is important to be just as critical of scientists as you would of any other authority. Although many scientists try to embody the ideals of science, these can be hard to uphold and you have no idea if any particular scientist does so. In addition, you would be naturally skeptical if research funded by a drug company supports the effectiveness of a drug manufactured by that same company. Similarly, if an organization with a particular socio-political agenda funds research that ends up supporting that agenda, you would also be skeptical of the findings. We don't always know the motivations and agenda associated with a piece of research, so it's generally best to be skeptical and evaluate the methods of all research closely.

As you practise scientific skepticism and learn more about scientific methods, you may increasingly question claims reported in the media. The website www.snopes.com can be helpful for evaluating urban legends and scams circulated online. You should also be vigilant for *pseudoscience*, which uses scientific terms to make claims look compelling and scientific, but actually falls short of using proper scientific methods. Pseudoscience ranges from astrology to some forms of self-help, with pseudoscientists often asking you to fork over money to improve your life, but without providing appropriate evidence to back up their claims. Detecting pseudoscience can be challenging, but there are some warning signs that suggest a claim might be pseudoscientific (see [Figure 1.2](#)).

Figure 1.2 Some warning signs of pseudoscience

- The claims are not falsifiable.
- If data are reported, the methodology producing these data is not scientific and the accuracy of the data is questionable.
- Supportive evidence is often anecdotal or relies heavily on authorities who are “so-called” experts in the area of interest. Genuine, peer-reviewed, scientific references are not cited.
- The claims ignore evidence that is contradictory.
- The claims are stated in scientific-sounding terminology and ideas.
- The claims tend to be vague, rationalize strongly held beliefs, and appeal to preconceived ideas and biases.
- The claims are never revised to account for new data.
- They propose a simple solution to your problems in exchange for money.



TRY IT OUT!

Let’s consider the claim that people can learn complicated material while sleeping (e.g., a second language). Before reading further, visit

www.sleeplearning.com. Adopt a scientifically skeptical mindset, use the list in [Figure 1.2](#), and try to detect evidence of pseudoscience.

If you've completed the *Try It Out!* task, then please read on. Note that the website explicitly claims that their work is based on science: "Latest research proves there is a link between sleep and learning." But this is a vague claim, that doesn't actually state that you can learn things in your sleep. Rather, it refers to the research showing that sleep can help you consolidate memories for experiences you have while waking (Rasch & Born, 2013). To their credit, this website paraphrases and quotes from some of these studies, but a major problem remains. There is absolutely no mention of evidence that listening to recordings during sleep can produce long-term learning, which is their central claim. Research that is somewhat relevant is used deceptively to increase the website's credibility. Other non-scientific forms of evidence is proffered, such as testimonials and anecdotes, but no science supports their claim. A general rule is to be highly skeptical when assertions are presented as scientific, but these are only supported by vague or improbable evidence. Especially when they are selling something that you wish was true, that you can learn without effort and study!



LO3 Goals of Scientific Research in Psychology

Four general *goals of scientific research* guide much of psychology and the behavioural sciences: (1) to *describe* behaviour, (2) to *predict* behaviour, (3) to determine *the causes* of behaviour, and (4) to understand or *explain* behaviour.

Describing Behaviour

The primary goal of scientific research in psychology is to describe behaviour and events. A good example of this can be found in the following Student Spotlight.

☆ Student Spotlight: Descriptive Research ☆

Adriana Knoll, working with Dr. Richard MacLennan at the University of Regina, wanted to find out just how prevalent major depressive disorder (MDD) is in Canada. In order to do so, they turned to data from the Canadian Community Health Survey—Mental Health, which polled a representative sample of Canadians. This allowed them to accurately

estimate just how common it is for people in Canada to experience MDD within a 12-month span, or within their lifetime, as well as whether this disorder is more likely to strike women or men. Based on this survey data, the 12-month prevalence rate for MDD was found to be about 5 percent, with the lifetime prevalence rate falling at about 11 percent. The odds-ratio for women versus men was 1.8, meaning that women were almost twice as likely to have MDD in their lifetime. Their research was published in the journal *Canadian Psychology* (Knoll & MacLennan, 2017).

Page 9Descriptive research like this is essential for understanding human psychology and human behaviour. Before we can predict or properly understand something, we first need to understand what that phenomenon is, what it looks like, when and how often it occurs, and for whom. Only after we have described a phenomenon well, its features and its prevalence, can we move forward to ask additional questions such as *why* it occurs. Are people with different personality traits more vulnerable to depression? Does living in an urban versus rural environment matter? What about living in different geographical zones that receive different levels of sunlight throughout the year? Are different forms of mental health treatment more effective for different kinds of people?

Predicting Behaviour

After we have identified and described a phenomenon or behaviour, a second goal of scientists is to be able to predict when that behaviour will occur and not occur. For example, if two events have been observed to be consistently related, it becomes possible to make predictions about when an event might occur and anticipate it. The following Student Spotlight illustrates this idea.

☆ Student Spotlight: Making Predictions ☆

Being a doctor is a stressful job, and many physicians suffer from burnout, a form of workplace exhaustion that can lead to a serious mental and physical collapse. Can we predict which doctors are more likely to experience burnout? Are doctors who tend to be more or less empathetic at

greater or less risk of burning out? Martin Lamothe and Dr. Serge Sultan at the Université de Montréal, along with some colleagues in Paris, examined this possibility by surveying almost 300 French doctors. The results were fascinating, and perhaps not what you might expect. Doctors who self-reported less ability to take the perspectives of others, and less empathy for others, were at higher risk of burnout. Their research now appears in the journal *BMC Family Practice* (Lamothe, Boujut, Zenasni, & Sultan, 2014).

However, we must be careful not to assume that such relationships imply causation, even though they help us make successful predictions. We will explore the appropriate conclusions that can be made from these correlational research designs in more detail in [Chapter 4](#).

Determining the Causes of Behaviour

A third goal of scientific research in psychology is to determine the causes of phenomena. Although we might accurately predict the occurrence of a behaviour, we might not have correctly identified its cause. For example, high school grades do not cause university grades, although they are related. There are likely various shared causes for both high school and university grades (e.g., motivation, study skills, conscientiousness, reading ability). In order to uncover these causes, we need to conduct focused research into these factors. Also, we often want to change behaviours (e.g., help students get better grades), and in order to do so we need to know the causes of those behaviours. The experimental method is what we use in science to help us to identify cause-and-effect relationships, when circumstances allow (see [Chapter 4](#)).

Criteria for Causal Claims

To make a causal claim, it is not enough to know that two events occur together. Cook and Campbell (1979) describe three criteria, drawn from the work of philosopher John Stuart Mill, that must be satisfied in order to identify a cause of some phenomenon. For example, consider research showing that multitasking on a laptop during a lesson predicts worse test scores, compared to just using that laptop to take notes (Sana, Weston, &

Cepeda, 2013). These researchers successfully made a claim that multitasking causes worse test scores by meeting the three criteria:Page 10

1. When the cause is present, the effect occurs; when the cause is not present, the effect does not occur. This is called *covariation of cause and effect*. Sana and colleagues showed that people who were multitasking on their laptop scored 55 percent correct on the test (on average), whereas people who were not multitasking scored higher (66 percent, on average).
2. The cause must precede the effect in time, known as *temporal precedence*. In this study, listening to the lesson while multitasking (or not) occurred before the comprehension test. In other words, it is just not possible that test scores somehow reached back in time and affected learning during multitasking.
3. Lastly, nothing other than the causal variable can be responsible for the observed effect. This is called *ruling out alternative explanations*. There should be no other plausible reasons for the relationship observed. Consider this hypothetical alternative explanation in this study: Suppose that everyone who multitasked also had no prior knowledge of the topic, whereas the people who did not multitask had all taken a course on the topic before. In this case, the difference in test scores could have an alternative explanation, other than the effect of multitasking: people who didn't multitask did better because they had prior knowledge of the topic. To avoid this, and many other alternative explanations, Sana and colleagues randomly assigned participants to either multitask or not during the experiment, thereby creating equivalent groups. The topic of causation, including the use of methods like random assignment to study it, will be discussed again in later chapters.

Explaining Behaviour

A final goal of scientific research in psychology is to explain *why* the events and behaviours occur. Consider the study in the Student Spotlight above that examined how empathy in doctors relates to burnout. What these

researchers found was that doctors who reported being higher in empathy, and more like to engage in perspective-taking, actually experienced less burnout from work. However, we still don't know exactly why this occurs. Is it because having empathy for one's patients reminds doctors of their purpose and the good that they manage to do? Or is it because doctors who engage in perspective-taking are actually more successful at diagnosing and treating patients, and so experience less stress associated with failure?

There are probably several other possible explanations for why more empathic doctors are less likely to experience burnout. However, additional research is needed to explore these different possible explanations. Such research often involves testing theories that are developed to explain particular relationships. Experiments and mediation analyses can both help us to explain behaviour (see [Chapter 4](#)).



Think about It!

What are some other possible explanations for why higher levels of empathy and perspective-taking in doctors are associated with a lower risk of burnout? Thinking of different possible explanations for an association is a great way to generate new research ideas, and to start thinking like a scientist!

Page 11Description, prediction, determining causes, and explanation are all closely connected. Determining cause and explaining behaviour are particularly closely related. Because it is very difficult to ever know all the causes for any behaviour, an explanation that appears satisfactory at first may turn out to be inadequate when other causes are identified. There is a certain amount of ambiguity in scientific inquiry. New research findings almost always invite new questions to be addressed by further research, and explanations of behaviour often must be discarded or revised as new evidence is gathered. Such ambiguity is part of the excitement and fun of science.



LO4 Basic and Applied Research

There are differences in the extent to which research is readily applicable to everyday contexts. The terms *basic* and *applied* research are often used to denote the ends of this continuum.

Basic Research

The four goals of science capture much of the focus of *basic research*, which attempts to answer fundamental questions about the nature of behaviour. Studies are often designed to develop and test theories about phenomena (see [Chapter 2](#)), such as cognition, emotion, motivation, learning, personality development, and social behaviour. Basic research often focuses on testing theories rather than developing a specific application. The results of basic research might eventually find applications, but these are not its primary purpose.

☆ Student Spotlight: Basic Research ☆

University of Toronto student Alyssa Sinclair, working with Dr. Morgan Barense, was curious whether a memory phenomenon observed in rats would also occur in humans. When viewing a surprising video for a second time, would a reminder of what this video is about midway through lead to interference, and thus some confusion between the memory for the old video and newly presented videos? In other words, do these reminders destabilize old memories and increase the likelihood that they can be modified by new experiences, creating false memories? In pursuing this question, they tested the predictions of two competing theories of memory. Although this research might have some useful applications down the road, the primary aim of this research is to better understand episodic memory, by testing different theories regarding how memory works. What they found was that this memory phenomenon does indeed occur in humans, and that these reminders can indeed destabilize memories. The researchers reported their results in the journal *Learning & Memory* (Sinclair & Barense, 2018).

Applied Research

Applied research is conducted to address practical problems in the real world and often to propose potential solutions. How rapidly this applied research is expected to be applied varies. Some applied research offers insight into problems or solutions, whereas other examples offer real tools to address those problems (Klatzky, 2009).Page 12

☆ Student Spotlight: Applied Research ☆

Jessica Lima, working with Dr. Martin M. Antony and his graduate student (now Dr.!) Hanna McCabe-Bennett, investigated a new way of treating those who have a fear of thunderstorms. These researchers from Ryerson University used virtual reality to provide a form of exposure therapy, and compared the efficacy of this approach with another established treatment: progressive muscle relaxation combined with psychoeducation. Fear of storms was measured both before and immediately after both kinds of treatment, as well as 30 days afterwards. After treatment, there was less fear of storms reported in both treatment groups, and this reduction in

symptoms was greater for those in the virtual reality group. Remarkably, these reductions in fear were still observed 30 days after treatment. In the case of this study, the practical real-world applications of this research are pretty obvious, aren't they? You can read find their study in *Behavioural & Cognitive Psychotherapy* (Lima, McCabe-Bennett, & Antony, 2018).

Careers in Applied Research

Some applied research is conducted at universities, and some occurs in settings such as large businesses, market research companies, government agencies, and public polling organizations. Some industry-based applied research is not published, but rather is used within the company itself or by the company's clients. Whether or not such results are published, however, they are used to help people make better decisions about problems. Your training in research methods could help you pursue a career in many of these fields.

A major area of applied research—and a growing career opportunity—is called [program evaluation](#). Program evaluation research tests the efficacy of social reforms and innovations that occur in government, education, the criminal justice system, industry, health care, and mental health institutions. (See also [Chapter 10](#).) Social programs are designed to achieve certain outcomes, and social scientists should evaluate each program to determine whether it is having its intended effect (Campbell, 1969). If it is not, alternative programs should be tried. The following Research Spotlight is an example involving program evaluation research for which program administrators collaborated with university researchers.

☆ Research Spotlight: Program Evaluation ☆

People involved in a syringe-exchange program for injection drug users in Vancouver collaborated with university researchers to examine the efficacy of their program (Hayashi, Wood, Wiebe, Qi, & Kerr, 2010). The researchers analyzed data collected via a survey of drug users and found that the program helped to reduce how often needles were reused, a highly risky behaviour that can have serious consequences (e.g., HIV infection).

Integrating Basic and Applied Research

Progress in science is dependent on a synergy between basic and applied research. Much applied research is guided by the theories and findings of basic research. For example, the creation of virtual reality depended on earlier psychology research into basic sensory and perception processes (e.g., perceiving visual depth). In turn, the findings obtained in applied settings often suggest modifications to existing theories and thereby inspire further basic research.

Some people, including legislators who control the budgets of government research-funding agencies, have demanded that all research be directly relevant to specific social issues. The problem with this attitude is that we can never predict the ultimate applications of basic research. The psychologist B. F. Skinner, for example, conducted basic research in the 1930s on operant conditioning, which carefully described the effects of reinforcement on such behaviours as bar-pressing by pigeons. Years later, this research led to many practical applications in therapy, education, and the workplace. But Skinner, and most in his era, had no idea this would be the case. Often, research with no apparent practical value can ultimately prove to be very useful. Coming to a better understanding of our world and the people within it is likely to have many benefits, most of them unknowable at the time of discovery. Because no one can predict the eventual impact of basic research, support for basic research is necessary to both advance science and benefit society.

Behavioural research is important in many fields and has many consequences for our daily life, including through changes in public policy. All researchers use scientific methods, whether they are interested in basic or applied questions. The themes and concepts we introduced in this chapter resonate throughout the remainder of this book. In [Chapter 2](#), we will consider how scientists find inspiration for research ideas and contribute to the body of knowledge by writing research reports. As you learn more about the scientific approach throughout this book, we invite you to engage in scientific skepticism and insist that claims be tested empirically.

Study Terms

Test yourself! Define and generate an example of each of these key terms.

- applied research (p. 11).
- authority (p. 4).
- basic research (p. 11).
- covariation of cause and effect (p. 10).
- empirical question (p. 6).
- empiricism (p. 4).
- falsifiable (p. 6).
- goals of scientific research (p. 8).
- intuition (p. 3).
- peer review (p. 6).
- program evaluation (p. 12).
- pseudoscience (p. 7).
- replicate (p. 6).
- ruling out alternative explanations (p. 10).
- scientific skepticism (p. 4).
- temporal precedence (p. 10).

Review Questions

Test yourself on this chapter's learning objectives. Can you answer each of these questions?

1. What are some of the benefits of learning about research methods?
2. Why is scientific skepticism useful in furthering our knowledge of behaviour? How does the scientific approach differ from other ways of gaining knowledge about behaviour?
3. Define and generate examples of the four goals of scientific research in psychology: (1) description of behaviour, (2) prediction of behaviour, (3) determination of the causes of behaviour, and (4) explanation of behaviour.
4. How does basic research differ from applied research?

Deepen Your Understanding

Develop your mastery of these concepts by considering these application questions. Compare your responses with those from other people in your study group.

1. Read several blog posts or editorials in a local newspaper and identify the sources used to support the assertions and conclusions. Did the writer use intuition, appeals to authority, scientific evidence, or a combination of these? Give specific examples.
2. Suppose you were assigned to participate in a debate about the following claim: Behavioural scientists should conduct only research that has immediate practical applications. Develop arguments that support and oppose the claim.
3. Suppose you were assigned to participate in a debate about the following claim: Knowledge of research methods is unnecessary for students who intend to pursue careers in clinical and counselling psychology. Develop a few arguments that support and oppose the claim. Next, use skills from [Chapter 2](#) and [Appendix D](#) to find the following article, and use it to evaluate your arguments:
 1. Lilienfeld, S. O., Ritschel, L. A., Lynn, S. J., Cautin, R. L., & Latzman, R. D. (2013). Why many clinical psychologists are resistant to evidence-based practice: Root causes and constructive remedies. *Clinical Psychology Review*, 33, 883–900.
4. A newspaper headline says that “Obesity Is More Common Outside Major Cities.” You read the article to discover that a researcher found that the rates of obesity are lower in cities such as Vancouver, Toronto, and Montreal than in less urban centres across the country. Based on this information, is it appropriate to infer cause and effect and

explanations of behaviour? Why or why not? Come back to this question after you have read the next few chapters. For more information, see:

1. Vanasse, A., Demers, M., Hemiari, A., & Courteau, J. (2006). Obesity in Canada: Where and how many? *International Journal of Obesity*, 30, 677–683.

Where to Start



©Bob Eastman/Getty Images

When you're just starting out, sometimes it helps to just go out on a limb and start exploring what's out there.

Learning Objectives

Keep these learning objectives in mind as you read to help you identify the most critical information in this chapter.

By the end of this chapter, you should be able to:

1. [LO1](#) Describe different sources of ideas for research, including questioning common assumptions, observations of the world, practical problems, scientific theories, and past research.
2. [LO2](#) Identify the two functions of a theory.
3. [LO3](#) Summarize what information is included in each section of a research article, including the abstract, introduction, method, results, discussion, and references.
4. [LO4](#) Compare and contrast different ways to find past research.
5. [LO5](#) Discuss how theories, research hypotheses, and predictions are related.

Page 16 SCIENCE IS MOTIVATED BY A NATURAL CURIOSITY ABOUT THE WORLD. People can be inspired to do research when their curiosity leads them to ask, “I wonder what would happen if . . .” or “I wonder why . . .” Research is a way to gather evidence that can shape our beliefs about what the answers to these questions might be. What are the sources of inspiration for such questions? How do you find out if other people and their past research are relevant to these questions? How do you develop your question so that a study can be designed to gather evidence regarding it? In this chapter, we will explore the first three steps in the process of conducting research, as depicted in [Figure 1.1](#): (1) generating an idea, (2) finding relevant past research, and (3) shaping your question into a testable hypothesis.



LO1 Where Do Research Ideas Come From?

Good research ideas can come from anywhere. Many people are capable of coming up with interesting ideas but find it difficult to say exactly how they come about. Sometimes ideas just seem to pop into our head and we can't say what inspired them. Other people have difficulty generating their own new ideas and might have to engage in a very deliberate process of thinking about what to study. When student researchers are just starting out, their research ideas might come from what they are studying in their courses, or from the professors they work with. Let's consider five broad sources of ideas for research: (1) common assumptions, (2) observation of the world around us, (3) practical problems, (4) scientific theories, and (5) past research.

Questioning Common Assumptions

One source of ideas that can be tested is the common assumptions that people have and that they use to explain the world. For example, researchers can question whether the *common-sense* or *folk-wisdom* beliefs held within a culture are actually true (Stanovich, 2013). Do "opposites attract" or do "birds of a feather flock together"? Is it friends or parents that

have the greatest influence on how children end up? Is a “picture worth a thousand words”? Asking questions such as these can lead to research on interpersonal relationships, child development, and the role of visual images in learning and memory, respectively.

Testing widely held assumptions can be valuable because these notions don’t always turn out to be correct. Research might also show that the real world is much more complicated than our assumptions would have us believe. For example, despite the common belief that opposites attract, decades of research has shown that people actually tend to be attracted to others who are similar to themselves (Montoya, Horton, & Kirchner, 2008; Newcomb, 1961). But things are also not so simple. One group of researchers led by Steven Heine (UBC) wondered if this finding would replicate in non-Western cultures (Heine, Foster, & Spina, 2009). What they found was that Japanese participants did not show this same preference for similar people over those who are different from them. Conducting research to test common assumptions often forces us to go beyond a common-sense theory of behaviour and examine more closely what actually occurs in the real world.

Observation of the World around Us

Page 17Simply making careful observations of what happens around us can lead us to develop intuitions about the world. But rather than accept these intuitions unthinkingly, we want to take on a scientifically skeptical mindset (see [Chapter 1](#)), pushing those intuitions to fuel research ideas. For example, have you ever set aside a certain amount of time to complete a project, only to find that it takes you much longer than you anticipated? Noting this experience could inspire future research on how and why people underestimate the time it takes to complete tasks. In fact, Roger Buehler (Wilfred Laurier) and his colleagues have conducted a series of experiments on this very topic (for a review, see Buehler, Griffin, & Ross, 2002). Their carefully designed studies showed that people are particularly likely to underestimate how long a project will take when it involves many steps (and therefore can be interrupted easily), rather than when the task can be completed in one sitting (Buehler, Peetz, & Griffin, 2010). Studying for exams and writing papers are exactly the kind of projects that take

many steps, so be careful to give yourself enough time to do them well! Adding a few extra hours, or days, to your first estimate might help you to correct for this bias.

Part-time jobs or volunteer positions can provide another rich source of material for scientific investigation. When he was a university student, psychologist Michael Lynn worked as a server in a restaurant, relying on tips from customers for income. The experience sparked an interest that fuelled his entire academic career (Crawford, 2000). For many years, Lynn has studied tipping behaviour in restaurants and hotels in the United States and in other countries (Lynn & Sturman, 2010). He has studied what increases tips—such as posture, touching, and phrases written on a cheque—and his research has had a major impact on the hotel and restaurant industry (Lynn & McCall, 2009). If you have ever worked in the service industry, you have undoubtedly formed many of your own thoughts about tipping behaviour. Lynn went one step further and took a scientific approach to testing his ideas. His research illustrates that taking a scientific approach to an everyday problem can lead to new discoveries with useful applications.

Keenly observing the world can help people take advantage of serendipity, or fortunate coincidence, to generate research ideas. Sometimes the most interesting discoveries are the result of accident or sheer luck. Ivan Pavlov is best known for discovering what is called *classical conditioning*: pairing a neutral stimulus (such as a tone) repeatedly with an unconditioned stimulus (food) to produce a reflex response (salivation), so that eventually the neutral stimulus presented alone will produce the response. Pavlov did not set out to discover classical conditioning. Instead, he was studying something else, the digestive system in dogs. One of his students happened to notice that the dogs began salivating before the actual feeding began, leading Pavlov to study the ways in which a stimulus preceding feeding could produce a salivation response (Windholz, 1997). Of course, accidental discoveries like this aren't purely accidental; they are far more likely when you approach the world with curiosity and an inquisitive eye.

Practical Problems

Recall from [Chapter 1](#) that the purpose of applied research is to address practical problems directly; the very existence of real-world problems can trigger an idea for a research project. Groups of city planners and citizens might survey bicycle riders to determine the most desirable route for a bike path, for example. On a larger scale, researchers have conducted research on major social and health issues, such as guiding public policy on graphic warning labels for cigarette packages in order to reduce smoking (Hammond et al., 2007). But drawing inspiration from real problems in the world is not limited to applied researchers. Those who tend to conduct basic research (see [Chapter 1](#)) may also draw inspiration from societal problems. Much of the basic research on memory, for example, was inspired by observations of the memory difficulties faced by people who had experienced some form of brain injury, such as a stroke. Most famously, observations of patient HM, who underwent brain surgery to help alleviate his epilepsy and experienced memory difficulties as a result, resulted in great advances in our understanding of basic memory processes (Squire, 2009).

Page 18



LO2 Theories

Much research in the behavioural sciences tests theories of behaviour. A [theory](#) is an organized system of logical ideas that are proposed to explain a particular phenomenon and its relationship to other phenomena (Fiske, 2004; Popper, 1968). In everyday conversation, people sometimes use the word theory to mean an idea that may or may not be true. This is not what

we mean by theory in science, so it's important not to confuse the two meanings. In fact, there is a scientific term that comes closer to that everyday meaning of theory, what scientists would call a *hypothesis*. We will explore the concept of hypothesis in more detail [later](#) in this chapter. For now, what is most important to remember is that a scientific theory is grounded in—and helps to explain—actual data from prior research. In explaining past observations, theories also specify predictions about possible future observations, known as hypotheses. These specific hypotheses can then be tested through further research, to help evaluate whether the broader theory is useful and likely to be true.

Theories serve two important functions in science. First, theories *organize and explain* a large number of previous observations. Often, these previous observations—data from past studies—are not as meaningful by themselves as when a theory can describe how they are all related to one another.

Theories can make the world more comprehensible by providing just a few abstract concepts around which many different observations can be organized. Consider how Charles Darwin's theory of evolution by natural selection organized and explained a large number of observations concerning the characteristics of animal species. This one theory helped to explain the appearance of so many different animals. In psychology, one theory of memory asserts that there are separate systems for working with information in the moment (working memory) and storing information for later (long-term memory) (Baddeley, 2003). This theory organizes many specific observations about learning and memory, including the different types of memory deficits that result from damage to different areas of the brain, and the rate at which a person forgets material he or she just read.

Second, theories help to *generate new knowledge* by pointing us in a direction where we can look to discover new aspects of behaviour. Theories help us generate new hypotheses about behaviour, which are then evaluated in future studies. If the studies support the hypotheses, then by extension, the theory is also supported. As a theory accumulates more and more supporting evidence, we become more confident that the theory might be correct as it enables us to explain many observations. Research may also reveal a weakness in a theory when part of it is not supported by evidence. When this happens, the theory can be modified to account for the new data.

Or in some cases, a researcher will develop an entirely new theory that accounts for this new data that a past theory, as well as past studies, could not explain. Science is fundamentally about basing our beliefs on systematic observation, so when new data arise, we should adjust our theories or abandon them for better ones. This process of building and amending our theories helps us to expand our knowledge of the world around us.

Theories are often modified when new research brings their limits into sharp relief. The necessity of modifying theories is illustrated by the theory of working memory and long-term memory mentioned previously.

Originally, the long-term memory system was described as a storehouse of permanent, fixed memories. However, research has now shown that memories are easily reconstructed and reinterpreted. In a classic study, participants watched a film of an automobile accident and were later asked to tell what they saw in the film (Loftus, 1979). Participants' memories were influenced by the way they were questioned. For example, participants who were asked whether they saw "the" broken headlight were more likely to answer yes than were participants who were asked whether they saw "a" broken headlight. Results such as these required the development of a more complex theory of how long-term memory operates.

Page 19

If multiple theories are equally successful at explaining the same phenomenon, the scientific principle of *parsimony* dictates that the *least* complex theory is the most desirable, because it is the easiest to falsify (Popper, 1968; see [Chapter 1](#)). Consider the case of two theories that differ in how many variables are used to explain the same phenomenon. Theory A is a complex theory with multiple variables, and Theory B is a simpler theory that focuses on just the variables that are necessary (and therefore is more parsimonious than Theory A). Now let's suppose that both theories really are wrong. It will be easier to show that all of Theory B is wrong than all of Theory A, because Theory B has fewer variables involved. Theory A, with more variables, will require more studies to completely prove it to be wrong. So the theory with the fewest links among variables is better because it is easier to entirely falsify than the theory with many links.

Past Research

Another rich source of ideas is past research. Becoming familiar with a body of research already published on a topic is perhaps the best way to generate ideas for new research. Virtually every study raises questions that can be addressed in subsequent research. The research may lead to an attempt to apply the findings in a different setting, to study the topic with a different population or in a different culture, or to use a different methodology to replicate the results. Recall from earlier that Steven Heine and colleagues (2009) found a result that replicated well in North America but did not hold in Japan. Exploring when and where phenomena arise means exploring different contexts and populations. Studies like these have led to important qualifications of past research about many psychological phenomena.

As you become familiar with the research literature on a topic, you may notice inconsistencies in these results: Some studies might find an effect, whereas others do not. Exploring why a phenomenon is observed inconsistently can provide a very valuable service to the field.

Alternatively, you may be interested in a highly replicated result but believe you have a better and alternative explanation for these results. Or you may want to propose a new theory to account for existing results and spur new ideas. Alternatively, you might use what you know about one research area (e.g., alcohol consumption) to inform what is known about another area (e.g., self-control; see Rawn & Vohs, 2011). These are all sorts of different ways that past research can inspire future research.

Let's look at an example of a study that was designed to address methodological flaws in previous research. The study examined a method of helping children who are diagnosed with autism. Childhood autism is characterized by multiple symptoms, including severe impairments in language and communication ability, and is now subsumed under the diagnosis of autism spectrum disorder (ASD; *Diagnostic and Statistical Manual of Mental Disorders*, Fifth Edition [DSM-5]). Parents and caregivers of children with ASD had been encouraged by a technique called *facilitated communication*. This technique purportedly allows a child to communicate with others by pressing keys on a keyboard, to show letters

and other symbols. A facilitator holds the child's hand to enable the child to determine which key to press. With this technique, many children with ASD seemed to communicate their thoughts and feelings and answer questions posed to them. Many people who see facilitated communication in action consider it to be a miraculous breakthrough, allowing the child to overcome the communication difficulties that often coincide with ASD.

Page 20

The original conclusion that facilitated communication is effective was based on comparing children's ability to communicate with and without the facilitator. The difference is impressive to many observers. Recall, however, that science invites skepticism, including examining all evidence carefully and asking whether claims are justified (see [Chapter 1](#)). In the case of facilitated communication, some researchers noted that the original study design failed to rule out a crucial alternative explanation: The facilitator may be unintentionally guiding the child's fingers to type meaningful sentences (Montee, Miltenberger, & Wittrock, 1995). In other words, the facilitator, and not the person with autism, was controlling the communication. They conducted a study to test this idea. In one condition, both the facilitator and the autistic child were shown pictures, and the child was asked to indicate what was shown in each picture by typing a response with the facilitator. In another condition, only the child saw the pictures. In a third condition, the child and facilitator were shown different pictures (but the facilitator was unaware of this fact). Consistent with the hypothesis that the facilitator was controlling the child's responses, the pictures were correctly identified only in the condition in which both saw the same pictures. Moreover, when the child and facilitator viewed different pictures, the child never made the correct response, and usually the picture the facilitator had seen was the one identified through facilitated communication. Despite replications of this demonstration that the technique does not actually allow children with ASD to communicate (e.g., Wegner, Fuller, & Sparrow, 2003), some practitioners have chosen to ignore this evidence and continue to believe that facilitated communication is an effective treatment (Mostert, 2010). This example, while illustrating the use of past research to generate new research questions, also reminds us that scientific skepticism is an important skill that helps us properly evaluate claims in everyday life.

Researchers draw inspiration for studies from any of the methods we have discussed, or even a combination of them. Questioning assumptions, making observations of the world, considering practical problems, examining current theories, and reviewing past research can all help us develop a research idea. In all cases, however, researchers need to find out what other scientists have already learned about the topic. Next, we consider how to find that past research literature.

How Do We Find Out What is Already Known?

Imagine you have observed something in the world around you and want to research it further. What comes next? It is time to find out what past research has already revealed about this topic. You might find a large body of past research, including well-developed theories, and this might lead you to revise your original question and ask something new. Or you might find that very little is known about the phenomenon driving your curiosity. Investigating past research helps the researcher to clarify the research idea and to design the study, ensuring that the study is making a new contribution to understanding behaviour. Reviewing past research can also help you in your everyday life when evaluating research reported in the media and when solving personal problems. In this section, we will explore what scientific papers look like, the different types of papers out there, and different ways to find them.

What to Expect in a Research Article

When a scientist has research results or a new theory to share with the scientific community, it is time to write up a report to submit for publication in a scientific journal (see [Figure 1.1](#) in Chapter 1). These reports usually share some common features. Becoming familiar with these features will help you to read and understand the research articles you find. As a new scientist, you might also be asked to practise writing up a research report that has these same features. For now, we will focus on what to look for while reading articles. See [Appendix A](#) for more information about how to write articles.
Page 21

Across all sciences, research articles that report the results of one study usually have six major sections: (1) an *abstract* that summarizes the entire report; (2) an *introduction* that explains the problem under investigation and any specific hypotheses being tested; (3) a *method* section that describes in detail the exact procedures used; (4) a *results* section that presents the findings; (5) a *discussion* section that might include speculation on the broader implications of the results, address potential alternative explanations, discuss reasons why a particular hypothesis might not have been supported by the data, and make suggestions for future research; and finally, (6) the *references* section that lists all the sources that

were cited in the article. In psychology, all citations throughout the main text of the article, and all references in the references section, follow the formatting rules found in the *Publication Manual of the American Psychological Association* (APA, 2010b). If you look to the back of this book, you will find that the references in the References section are formatted using these rules. This manual also provides rules for formatting the other parts of the article, such as how to prepare tables and how to present statistics. Psychologists follow a common set of rules for formatting to help communicate with each other efficiently. Because of these rules, all scientists know where to find specific information (e.g., where to find the procedures used: in the methods section). Below, we elaborate a bit on each of the different sections. Note that some research reports have more than one study, and so they include methods and results sections from each study. All these results are later synthesized later in a *general discussion* section. Take a look at [Figure 2.1](#) for an overview of all the sections, then read on for detailed descriptions of each.

Figure 2.1 Major sections of a research article

Abstract	Brief summary of the article	Tends to start broadly (with a statement of the topic) and narrow toward study method
Introduction	Outline the problem, tie to past research, point to question and method	
Methodology	Detailed description of study design	
Results	Objective report of study results	
Discussion	Interpretation of study results	Tends to recap results and then provide more general information
References	List of all works cited	





LO3 Abstract

The *abstract* is a summary of the research report that is found right at the beginning. It typically runs no more than 120 words in length, although the word limit can vary by journal. It includes information about the hypothesis, the procedure, and the broad pattern of results. Generally, little information is abstracted from the discussion section of the paper. Abstracts help readers quickly get a sense of a paper to decide if it is relevant to their interests and something that they may wish to read about in more detail. Simply reading an abstract doesn't tell you the whole story, however, and relying too heavily on this very brief summary can lead to a mistaken impression. It is important to read the entire article in order to adopt a properly skeptical scientific mindset. Being a scientist means being skeptical of everything we encounter, even other scientific reports. *Read the abstract to decide whether the article could help you learn about your research topic.*

Introduction

In the *introduction*, the researcher outlines the problem that has been investigated. Past research and theories relevant to the problem are described, a gap in the existing knowledge is identified, and the current study is introduced as an attempt to fill this gap in knowledge. The researcher will often end this section by stating the hypothesis, or by declaring the study to be exploratory yet guided by particular research questions. In other words, the investigator introduces the research project by building a logical case that justifies why this study, and the expected results, will make an important contribution to understanding behaviour. *Read the introduction to find out the purpose of the study, the past research and theories relevant to the study, and the hypothesis.*

Method

The [method](#) section provides information about exactly how the study was conducted, including any details necessary for the reader to replicate (repeat) the study. It is often divided into subsections, with the number of subsections determined by the author and dependent on the complexity of the study design. One subsection always describes the characteristics of the participants who contributed the data. These individuals may also be referred to as subjects, or respondents when surveys are employed (see [Chapter 7](#)), and sometimes as informants when people report on other people (APA, 2010b; Vazire, 2006). It is important to provide details of who contributed the data, so we can place these results in context and consider to what similar populations these results might generalize (for a discussion of [generalization](#), see [Chapter 14](#)). How many participants identified themselves as male, or female, or transgender, or non-binary, for example? What was the average age? How many participants were included in total? If the study used human participants, how were they recruited for the study? Was the data from any participants excluded from the analysis, and if so why? If the study used non-human animals, what species and genetic strain was used? Another subsection typically details the procedure used in the study, including the stimulus materials presented and the measures employed. It is important that no potentially crucial detail be omitted while describing stimulus materials, the way behaviour was recorded, and so on. Omitting important details prevents readers from properly evaluating the study. *Read the method section to find out characteristics of the participants, what they were asked to do, what materials were used, and the overall study design.*

Results

In the [results](#) section, the researcher presents the findings, which have typically arisen from a statistical analysis of the data collected. The presentation of these results takes various forms. Findings can be described in a narrative form, in which the result is simply described in words. An example of this might be, “The names of the animals that were also rated as cuter were more likely to be remembered one week after they were learned.” Researchers often try to avoid presenting too much interpretation of the results at this point, with extensive comments about what the results mean typically reserved for the discussion section. Research findings are also described in the form of statistics, which reflects the analyses that were conducted to test the hypothesis. These might look something like this, “ $r (283) = .21, p = .009; 95\% \text{ CI } [.08, .36]$.” (Don’t worry if that doesn’t make sense to you right now; you’ll learn about this later.) These

statistics can also be in tables and/or presented in the form of graphs or figures.

Page 23

At the outset, the statistical terminology of the results section may appear intimidating. Think of statistics as a tool the researcher uses to evaluate the outcomes of the study. If you want to become a scientist, you will learn these tools eventually. Right now, you can begin to build your ability to understand statistics by learning the logic behind them, even before you learn the specific calculations. [Chapters 12](#) and [13](#) and [Appendix B](#) provide a brief introduction to the powerful tools of statistics. *Read the results section for sentences, statistics, tables, and graphs that summarize the pattern of findings. This section will become easier to read as you increase your knowledge of statistics.*

Discussion

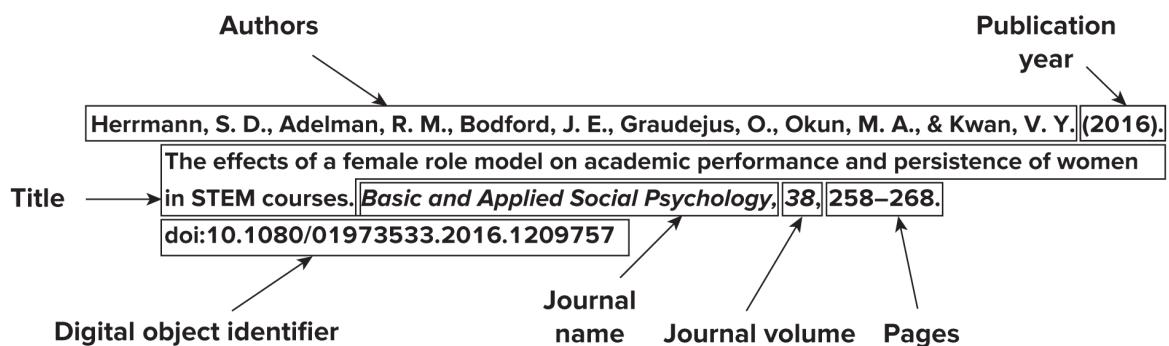
In the [*discussion*](#) section, the researcher reviews the current study from various perspectives. Do the results support all the hypotheses, some of them, or none of them? The author should try to discuss possible explanations for the results and discuss why one explanation might be more likely than others. If the hypotheses have not been supported, or have received only partial support, the author should suggest potential reasons. What might have been wrong with the methodology, the hypothesis, or both? There is nothing wrong with admitting that you realized a flaw in your study only after the data have been collected and analyzed. All studies have flaws and no study is perfect: It's very important to be upfront about the weaknesses and limitations of your particular study so that readers can probably evaluate your findings. In this section, the researcher also discusses how the results compare with past research on the topic. Lastly, this section frequently includes suggestions for future research on the topic, and possibly some practical applications. *Read the discussion section for conclusions about the hypothesis, the study's strengths and limitations, and contributions it makes to understanding the topic.*

References

Throughout any scientific article and many textbooks (including this one), you will notice [*citations*](#): names and dates within parentheses that often appear at the end of sentences. These citations are a short form of the full reference, which is listed alphabetically (by the last name of the article's first author) in the [*references*](#) section at the end of the report. Citations are a bit like links on a website, which connect readers to additional material. Sometimes, links on a site

function just as citations, guiding readers to the original source of material being referred to on the post. Citations in academic writing signal to readers that the idea or result described in that sentence was stated or found by the people cited, and *not* the current author. In some disciplines it is common to include direct quotations along with citations in research articles, but writers in psychology and other sciences tend to paraphrase others' work instead (Madigan, Johnson, & Linton, 1995). *See what we just did there?* We signalled to you that Madigan and colleagues found that psychologists tend to paraphrase more than other disciplines. It's not our finding. We are using their finding to make a point, so we include a citation to their work to indicate as such. You can go to the References list in this book and find the full reference to this article by Madigan and colleagues, which you can use to find the full text of that article. [Figure 2.2](#) shows a breakdown of the elements in a typical reference for a journal article. If you read this article, then you can decide for yourself if we paraphrased their findings correctly—all part of adopting a skeptical scientific mindset. *Read the references section to look up citations you noticed in earlier sections. Use this information to find research articles that will help you learn more about what is known about your topic.*Page 24

Figure 2.2 Anatomy of a standard reference



Other Types of Articles: Literature Reviews and Meta-analyses

Articles that review and summarize research in a particular area may not closely follow the six-section format described above. Depending on the way the research is summarized, this type of article might be called a [literature review](#) (if it uses narrative techniques to summarize the research using words only) or a [meta-analysis](#) (if it uses statistical techniques to re-analyze past data; see [Chapter 14](#)). These articles might also propose new theories to better explain existing

research. Review articles will have an abstract, introduction, discussion, and references section, but the sections between the introduction and discussion will vary. The following Student Spotlights provide examples of a literature review and a meta-analysis.

☆ Student Spotlight: Literature Review ☆

Tessa Grant, working with Dr. Elizabeth Kelley at Queen's University, reviewed research relevant to whether a diagnosis of autism spectrum disorder (ASD) should be considered when making judgments of criminal responsibility. In order to do so, they summarized and integrated past research on perspective-taking and moral-reasoning abilities in this population. Based on their survey of the relevant research, they recommend that diagnoses of ASD should be considered in Canadian courts, with respect to criminal responsibility. They also discuss the limitations of available research in this area and suggested some ways that future research could improve our understanding of this topic. You can find their review article in the journal *Canadian Psychology* (Grant, Furlano, Hall, & Kelley, 2018).

☆ Student Spotlight: Meta-analysis ☆

While at the University of Manitoba, Michelle Ward completed a meta-analysis of past studies examining whether parent-child interaction therapy benefits children with various forms of disruptive behaviour disorder. Working under the supervision of Dr. Jennifer Theule, they found 12 different studies investigating this possibility and statistically aggregated these past results. In doing so, they found that this form of therapy was very effective in helping children with disruptive behaviour disorder, regardless of the child's gender. This meta-analysis appeared in *Child & Youth Care Forum* (Ward, Theule, & Cheung, 2016).

Page 25

Reading Articles

Reading articles is a great way to become familiar with the way information is presented in scientific reports. (Consider starting with the entire article that appears in [Appendix A](#). It is in manuscript form, formatted the way a psychologist would submit it for potential publication, with notations to help you create your own manuscript in APA style.) As you read more and more articles,

you will develop ways of efficiently processing the information in them. You do not need to read every section in order. It is usually best to read the abstract first, and then skim the article to decide whether you can use the information provided. Sometimes it is appropriate to skip to the section you need most. For example, when searching for ways that other researchers have studied creativity, you might head straight to the methods section. Then go back and read the article carefully. Note the hypotheses and theories presented in the introduction, write down anything that seems crucial or problematic in the method, and read the results in view of the material in the introduction. Be critical when you read an article. As you read more research on a topic, you will become more familiar with the variables being studied, the methods commonly used to study those variables, the important theoretical issues being considered, and the problems that need to be addressed by future research. In addition, once you learn more about statistics, you will find yourself reading results sections more critically, considering whether they conducted their analyses in an appropriate way. While reading articles, you just might find yourself generating your own research ideas and planning your own studies.



LO4 Where Are These Articles Published? An Orientation to Journals and Finding Articles

Once a researcher has written a paper, it is time to submit the article to a scholarly scientific journal to be considered for publication. Scholarly journals are special publications that include articles written by scientists that are intended for other scientists. If the journal is peer-reviewed, then the journal's editor will solicit reviews from other scientists in the same field and will use these reviews

to decide whether the report is to be accepted for publication. This is the process of peer review introduced in [Chapter 1](#). (Note that all peer-reviewed journals are scholarly publications [intended to be read by other scholars and scientists], but *not all* scholarly publications are peer-reviewed.) Each journal receives many more submissions for consideration than it can publish, and many of these submissions might not meet various criteria for the journal (e.g., relevant research topic, study quality). As a result, most papers submitted to a journal are rejected. This does not mean, however, that every article published in a peer-reviewed journal is of high quality and without limitations. It is still very important to approach published peer-reviewed articles with a critical eye. Articles that are accepted are published in print about a year later; however, many journals now publish a version of the article online shortly after it is accepted. Other journals are published entirely online and never in print (e.g., *PLoS ONE*). Once accepted for publication by a journal, these research reports are considered primary sources in our discipline.

There are very many scientific journals. Some are published monthly, others quarterly or annually. Some journals publish articles that cover many different topics across a discipline (e.g., *Psychological Science*), whereas others are more focused on a particular subfield in the discipline (e.g., *Cognition & Emotion*). Yet other journals publish only literature reviews and meta-analyses (e.g., *Annual Review of Psychology*). [Table 2.1](#) lists just a few of the journals for several different areas of psychology.

Page 26

Table 2.1 Some Major and Specialized Journals in Psychology

General

American Psychologist

Collabra: Psychology

Canadian Journal of Behavioural Science

Journal of Experimental Psychology: General

Canadian Psychology

Psychological Science

Literature reviews and meta-analyses

Current Directions in Psychological Science

Psychological Review

Perspectives in Psychological Science

Psychological Science in the Public Interest

Revue Quebecoise de Psychologie

Cognition and behavioural neuroscience

General

Behavioral Neuroscience

Journal of Experimental Psychology: Animal Learning and Cognition

Canadian Journal of Experimental Psychology

Journal of Experimental Psychology: Learning, Memory, and Cognition

Canadian Journal of Neurological Sciences

Neuropsychology

Cognition

Clinical and counselling

Behavior Research and Therapy

Journal of Abnormal Psychology

Journal of Abnormal Child Psychology

Journal of Consulting and Clinical Psychology

Developmental

Child Development

Developmental Review

Journal of Experimental Child Psychology

Developmental Psychology

Personality and social psychology

Comprehensive Results in Social Psychology

Journal of Personality and Social Psychology

Journal of Experimental Social Psychology

Personality and Social Psychology Bulletin

Journal of Personality

Social Psychological and Personality Science

Applied and interdisciplinary

Journal of Applied Psychology

Health Psychology

Journal of Educational Psychology

Journal of Consumer Psychology

Cyberpsychology, Behavior, and Social Networking

Journal of Experimental Psychology: Applied

Evaluation and Program Planning

Media Psychology

Psychology, Public Policy, and Law

Sexuality, gender, culture, and family studies

General

Journal of Cross-Cultural Psychology

Journal of Sex Research

Journal of Family Psychology

Psychology of Men and Masculinity

Journal of Marital and Family Therapy

Psychology of Women Quarterly

There are so many journals it would be impossible to list them all—let alone read them all! This is why we need search engines and databases to help us find the articles that are relevant to our interests. Next we offer some tips for finding articles published in scholarly journals. For further information, we encourage you to consult your institution’s library resources (e.g., your library’s website, librarians, and any workshops the library may offer), as well as more-detailed guides about searching for articles and how to write scholarly papers (e.g., Reed & Baxter, 2003; Rosnow & Rosnow, 2012). Some institutions have subject librarians, who specialize in helping students and faculty working in a particular field. These librarians can be a valuable resource for how to find the articles you need for a research project. Check out what resources are available at your institution to help you.

Page 27
Although you can access physical copies of printed journals by visiting your institution’s library, most articles are available online as PDF or HTML documents. Finding them—and finding the right articles—can be more challenging than it seems at first, especially if you are new to a topic. It takes some time to learn the precise terminology used in each field, and it might take a few hours to find just the right keywords to search with that will reveal the most relevant articles. We will consider a few different ways of searching for articles online, and how best to use each of them. [Table 2.2](#) lists each method of searching and highlights some pros and cons of each type of search.

Table 2.2 Comparing Library Databases with Internet Searches

Source	Useful for...	Cautions and Caveats
---------------	----------------------	-----------------------------

Source	Useful for...	Cautions and Caveats
Library databases (e.g., <i>PsycINFO</i> , <i>Web of Science</i>)	<p>Finding specific articles</p> <p>Setting advanced search parameters (e.g., keywords, peer-reviewed articles only, author name)</p> <p>Reliable cited reference searches</p> <p>Complete, searchable bibliographic details</p> <p>Trusting that the database is monitored for accuracy</p> <p>Accessing many full-text articles when connected to campus library</p> <p>Learning disciplinary jargon by using thesaurus-suggested terms as well as keyword lists from useful articles</p>	<p>For full access, log in on campus or consult your library resources for how to log in while off campus.</p> <p>Knowing the important terminology in your topic is helpful for choosing specific search terms.</p>
Internet search	<p>Exploring a topic very broadly, including blog posts, magazine articles, videos, etc.</p> <p>Finding researchers' websites for their full publication records and results</p>	<p>Cannot set advanced search limits (e.g., to limit to peer-reviewed articles)</p> <p>Can be difficult to find peer-reviewed journal articles amid broad, general search results</p> <p>Not monitored for accuracy</p>

Source	Useful for...	Cautions and Caveats
Scholar.google.ca	<p>Exploring topic fairly broadly within books, dissertations, peer-reviewed journal articles, patents, etc.</p> <p>Setting basic search limits (e.g., publication date)</p> <p>Accessing some full-text articles when connected to campus library</p> <p>Finding some references to journal articles for <i>PsycINFO</i> search</p> <p>Learning some important terminology for a new area of study</p> <p>Finding sources for multidisciplinary topics</p>	<p>For more full-text links, log in on campus or consult your library resources for how to log in while off campus. Results include all disciplines (e.g., search for “power” yields entries from sociology, political science, psychology, statistics, physics, as well as all articles with an author whose last name is “Power”). Cannot set advanced search limits (e.g., for peer-reviewed articles)</p> <p>Can be difficult to find peer-reviewed journal articles amid dissertations and books</p> <p>Citation counts are unverified</p> <p>Not monitored for accuracy</p>
Wikipedia	<p>Learning some important terminology for a new area of study</p> <p>Finding some references to journal articles for <i>PsycINFO</i> search</p>	<p>Not monitored for accuracy</p> <p>Entries often have greatly limited or absent references</p> <p>Not considered an acceptable source for most research papers and assignments</p>

Page 28

PsycINFO

There are many databases that cover a wide range of academic disciplines. The American Psychological Association’s searchable database system is called [PsycINFO](#), and it includes coverage of publications in psychology and related fields from the 1800s to the present. This is considered the best, most authoritative source for finding psychology articles. The exact procedures you

will use to search *PsycINFO* will depend on your university’s library system; you should consult your library to find out how to access *PsycINFO*. (This is true for many databases, such as *PubMed*, which is a great resource for biological psychology and biomedical literature.) Once you have access to *PsycINFO*, you enter search terms and any other parameters you like (e.g., specifying peer-reviewed journals only) to obtain a list of abstracts that meet your search terms. *PsycINFO* is useful because it allows us to limit search results, improving the likelihood of finding exactly the articles we need. For example, we have the option to limit results to a particular date range or just one specific author.



TRY IT OUT!

Examine [Appendix D](#) for the steps to conduct a *PsycINFO* search. Try following along using *PsycINFO* through your institution’s library. Your library’s website should include instructions for how to access these resources from off campus, if need be.

Web of Science

Another excellent resource is the *Web of Science*, which can be accessed directly or through the *Web of Knowledge* database. Like *PsycINFO*, [*Web of Science*](#) allows you to search for articles using citation information, such as the name of the author or the article title. It draws articles from many disciplines, including biology, chemistry, and pharmacology, as well as psychology, sociology, and criminal justice. One strength of *Web of Science* is its ability to conduct a “cited reference search.” This search allows you to find the other articles that have cited a particular article, perhaps one that is very relevant to your topic. Doing a cited reference search involves first identifying a key article on your topic. You can then search for other articles that have cited this key article. This search will give you a list of other articles that may be relevant to your topic. (Note that *PsycINFO* may offer a cited reference search service too, depending on your library’s agreements.) To provide an example of this cited reference search process, consider the following article:

- Ross, M., Xun, W. Q. E., & Wilson, A. E. (2002). Language and the bicultural self. *Personality and Social Psychology Bulletin*, 28, 1040–1050.
doi:10.1177/01461672022811003

An article search using the *Web of Science* resulted in 171 articles that have cited this paper since it was published in 2002. Here is one of them:

- Cheung, B. Y., Chudek, M., & Heine, S. J. (2011). Evidence for a sensitive period for acculturation: Younger immigrants report acculturating at a faster rate. *Psychological Science*, 22, 147–152. doi:10.1177/0956797610394661

After becoming familiar with this article or others on the list, you might use it as a new key article for further searches. Another important use of the cited reference search is to find all the articles that have used a particular measure. Starting with the key publication that introduces a particular measure, a cited reference search should help you identify other articles that have used this same measure (along with other articles that are mentioning it for other reasons). This can be very helpful if your research topic is concerned with measurement (i.e., psychometric in nature). It is also possible in *Web of Science* to specify a key researcher in order to find all articles written by or citing a particular researcher.

Other Library Databases

Your library may or may not have access to *PsycINFO* or *Web of Science*. The number of information databases that a library may purchase today is enormous, and budget considerations determine which ones are available to you. You will need to take advantage of any instructional materials that your library provides to help you learn how to best search for information available through your library. Other major databases include *Academic Search Complete*, *Sociological Abstracts*, *MEDLINE*, *PubMed*, and *ERIC (Educational Resources Information Center)*. In addition, services such as *PsycEXTRA*, *Canadian Newsstand Complete*, and *Access World News* allow you to search general media resources such as newspapers.

Some of the resources available provide the full text of the articles in the database, whereas others provide only the abstract or citation. *PsycINFO* typically links to full-text access for most journals, depending on your library's subscriptions. If the full text of the article is not available via a link, you may be able to obtain it by contacting one of the authors or by requesting it from another library. A reference librarian can help you to get access to an article from another library.

The Internet

There is a wealth of material that is freely available on the Internet. Search engines such as Google allow you to search to help you find websites devoted to your topic, including articles people have posted, book reviews, and online discussions. Although it is incredibly easy to search the Internet, you can improve the quality of your searches by learning (1) the differences in how each search engine finds and stores information; (2) advanced search rules, including how to make searches more narrow and how to find exact phrases; and, perhaps most importantly, (3) ways to critically evaluate the quality of the information that you find. It is a good idea to keep careful records of your searches, noting what search engine and search terms you used, the dates of your search, and the exact location of any websites that you think will be useful. This information will be useful when you provide citations and references for the information that you use while preparing papers, and may help you avoid unintentional plagiarism (see [Chapter 3](#)).

When using the Internet for research and in everyday life, it is essential to critically evaluate the quality of the information that you find. Your university's library and a variety of other websites all have information on how to critically evaluate what you find on the Internet. But for the most part, the safest source of information for your academic writing is going to be primary sources (i.e., scholarly articles from peer-reviewed journals). Some of the most important things to look for when considering information on the Internet are listed here:

- Are credible references (e.g., peer-reviewed articles) provided for factual information? Have you looked at these references yourself to judge their suitability and credibility?
- Is the site associated with a major educational institution or research organization? A site sponsored by a single individual or an organization with a clear bias should be viewed with extra skepticism.
- Is information provided about the people who are responsible for the site? Can you check the credentials of these people? Is contact information provided for the authors, to enable you to verify the credibility of the site?
- Is the information current?
- Do links from the site lead to legitimate organizations?

Google Scholar

Google has developed a specialized scholarly search engine called Google Scholar (<http://scholar.google.ca>). When you conduct a Google Scholar search, you find papers and books from scholarly journals, universities, and academic book publishers. This can be a useful addition to searching *PsycINFO*, particularly for topics that span across different disciplines (e.g., artificial intelligence spans psychology and computer science). However, one of the primary disadvantages is that Google Scholar does not allow you to narrow your search as precisely as *PsycINFO*. For example, you cannot limit your search to peer-reviewed research, but instead you have to search through a listing that includes books, dissertations, or book reviews that are typically less helpful for research projects than peer-reviewed articles. Another disadvantage of searching using Google Scholar is that the full text of primary source articles is sometimes unavailable or available only for pay. Be careful not to pay for articles your library provides for free! If you access Google Scholar while on campus, or while logged into your library account, you might be able to access full-text articles through Google Scholar using your library's subscriptions.

Wikipedia

The Internet is overflowing with content, and this content can be generated by anyone, with or without expertise. Wikis such as Wikipedia invite any user to adapt its content. Although the ultimate goal of Wikipedia is to be a trustworthy source of information (Wikipedia, 2011a), it is *not intended nor considered a credible source for academic research* (Harris & Cameron, 2010; Wikipedia, 2011b; Young, 2006). Sites like Wikipedia, as well as encyclopedias, can be useful starting points to learn about terminology or a key idea in your topic area (e.g., see <http://en.wikipedia.org/wiki/Portal:Psychology>). However, Wikipedia is simply that: only a starting point. For example, you might be interested in how people perceive depth. If you choose to search Wikipedia for “depth perception,” you can see that there are both monocular and binocular cues to distance. Scrolling down to the bottom of the page, there is a list of references that you can use as a starting point in your *PsycINFO* or *Web of Science* searches to investigate these types of cues. If there are no academic journal articles listed (and especially if there are no references at all), this is a clue to be even more skeptical than usual about the quality of what appears in the Wikipedia entry.

Comparing and Evaluating Search Results

You may find the most effective strategy to find the past research you need is to combine results from various sources, such as Google Scholar and *PsycINFO*. As

you use different types of searches, you will learn more about the topic, which will make it easier to use search tools more efficiently. As you search, note the keywords you see most often associated with the journal articles that are most relevant to your topic, and use those words in your subsequent searches.



LO5 Developing Hypotheses and Predictions

So you have come up with a research idea and have refined it after incorporating the past research you found, using the [above](#) search strategies. Now it might be time to create a hypothesis. After reviewing the past literature on a topic, researchers will know whether enough is already known about it to formulate a hypothesis. A [*research hypothesis*](#) (often just called a hypothesis) is a statement about a phenomenon that may or may not be true, informed by past research or derived from a broader theory, and which requires further evidence to support or refute it. If very little or no research has been conducted previously on a topic, then there might not be enough existing evidence to formulate a strong hypothesis. In this situation, a researcher might choose to conduct purely exploratory research that is not aimed to test a particular hypothesis, but instead simply to see what emerges as a result of the study. If a hypothesis is being tested by a study, however, then the data must be evaluated in terms of whether it provides evidence that is consistent or inconsistent with the hypothesis. See the Student Spotlight below for an example of how research questions relate to hypotheses.

Page 31

☆ Student Spotlight: Questions & Hypotheses ☆

Dan Tao, like many other senior undergraduate students, was thinking about career choice. More specifically, Dan was curious about how culture influences career decisions for those who have a foot in two different cultures, known as biculturals. This raises several interesting questions, such as “Does identifying with Canadian culture mean deciding on a career based on an analytic-informational approach?” and “Does identifying with Chinese culture mean choosing a career based on a more normative approach rooted in the expectations of family?” These questions were formalized into the following hypothesis: “The cultural orientation of Chinese Canadians should influence one’s approach to career decisions, and this is a result of orientations toward the self, family, and different life goals” (Tao, Zhang, Lou, & Lalonde, 2018).



TRY IT OUT!

Wait a minute, what did Dan Tao and collaborators discover about cultural identification and career choice? You might have noticed that we have deliberately left out the results of their study. This is to encourage you to track down the paper and find out for yourself. Using the skills you just learned about finding articles, can you find the abstract for this article? The entire article? Try to exercise these skills by doing the same for all Student Spotlights throughout this book.



Think about It!

Imagine what it would be like to write an exam in a crowded lecture hall, squeezed in elbow-to-elbow. Now imagine what it would be like to write an exam in a room separated from other people by at least a metre in all directions. Do you think the crowding would have an effect on your performance? Based on this potential experience, now imagine that you wanted to conduct an experiment. What might be your research hypothesis? Can you state it clearly and concisely?

Once a research hypothesis is stated, it is time to design a specific study to test it. For example, if you wanted to test your hypothesis from the *Think About It!* box above, you might ask participants to complete spatial rotation or memory tasks in either a crowded or uncrowded room. Performance on these tasks could then be compared to assess whether crowding affected performance. When designing the study, the researcher translates the more general hypothesis into a very specific *prediction* concerning the outcome of this particular study (see [Chapter 4](#)). Importantly, predictions are tied to the methods of a particular study. For example, your general hypothesis from the *Think About It!* example might have been something like:

- A crowded environment results in worse performance on cognitive tasks compared to an uncrowded environment.

In comparison, your specific prediction for this study might be more like:

- Participants in the uncrowded condition will perform better on a delayed recall task than participants in the crowded condition.

Note the differences between the general hypothesis and the specific prediction stated below. For our prediction, we have taken “cognitive tasks” and made them more concrete, specifying that in the context of this study we are using a “delayed recall task.” Another subtle shift you might notice is that predictions are often stated in the future tense (e.g., “will perform”), because these predictions appear in the Introduction section and refer to the study that is about to be described.

If the results of the study are consistent with the prediction, the more general research hypothesis is also supported. If the results are not consistent with the prediction, the researcher will either reject the hypothesis (and conclude that crowding does not lead to poor performance) or conduct further research, testing the hypothesis using different methods (e.g., different cognitive tasks). When the results of a study are consistent with a prediction, then we have evidence that is consistent with the hypothesis being correct, and so it is only *supported* by the evidence: It is not *proven* to be correct or incorrect. Science is an ongoing process of gathering evidence to inform our beliefs about what might or might not be true: It is not an exercise in proving things to be correct or incorrect.

Scientists are skeptical by nature, and so one study finding evidence consistent with a hypothesis won't be convincing on its own. Researchers will study the same hypothesis using a variety of methods in order to gather as much evidence as they can. If it is the case that each time this hypothesis is investigated it is supported by the evidence, then we might become more and more confident that the hypothesis is likely to reflect the truth. However, a good scientist is always willing to adjust one's belief in a particular hypothesis in the face of new and compelling evidence.

At this point, it is important to remember a key characteristic of all scientific hypotheses, which we discussed [earlier](#) in this chapter and in [Chapter 1](#): falsifiability (Popper, 1968). This means that data could show that a hypothesis is false, *if* in fact it is false. Consider the hypothesis from the Student Spotlight above, which asks whether and how cultural orientation influences career choices for Chinese Canadians. It is certainly possible to measure and compare the level of cultural orientation within bicultural individuals along with their approach to making a career choice. If the degree to which people identify with their Chinese heritage doesn't predict their orientation toward career decision-making, then this hypothesis would be falsified. Also, if cultural orientation does predict how people make career choices, the hypothesis is still capable of being falsified—but instead the data have supported it. In contrast, consider a hypothesis that is unfalsifiable: People have an invisible aura that changes colours depending on the person's health or illness. If the aura is invisible, how can it be measured at all? What kind of data could ever support or disprove this hypothesis? Because scientists ascribe to empiricism (see [Chapter 1](#)), hypotheses are scientifically meaningful only if they can be falsified through scientific means (i.e., through systematic observation). This aura hypothesis cannot be falsified using objective data obtained through observation, therefore it is a scientifically meaningless hypothesis.

In this chapter, we have elaborated on the first three steps of the research process: (1) generating a research idea, (2) learning about past relevant research, and (3) formulating a hypothesis (see [Figure 1.1](#)). We considered different sources of inspiration for research ideas, how to find past research and what research articles look like, and how to transform an idea into a testable research hypothesis leading to a specific study prediction. The next

steps involve designing a study to test that hypothesis, collecting data, and evaluating whether the data support that hypothesis. We will explore many options for study designs throughout the rest of this book. But first, in [Chapter 3](#), we consider the importance of ethics at all stages of the research design and reporting process.



Illustrative Article: Laptops in Class

Ravizza, Uitvlugt, and Fenn (2017) studied computer use in college and university classrooms by having students in an introductory psychology class log into a proxy server that monitored all online activity during class. They were specifically interested in the relationship between Internet use and classroom performance.

First, search for the article in *PsycINFO*, acquire a full-text version, and read the article:

- Ravizza, S. M., Uitvlugt, M. G., & Fenn, K. M. (2017). Logged in and zoned out: How laptop Internet use relates to classroom learning. *Psychological Science*, 28, 171–180. doi:10.1177/0956797616677314

After you have a full copy of the article:

1. Read the introduction:
 1. Summarize the authors' purpose for conducting the study; note that their purpose is supported by other references.
 2. Identify the research question or questions that are being explored by this study. In some studies, research questions are clearly labelled as such.
 3. Identify the hypotheses that are being tested by this study. As with research questions, in some studies, hypotheses are clearly labelled as such.

2. Read the Methods section:

1. How many participants did they invite to participate? How many participants did they collect data from? What was the breakdown of the students' class-rank?
2. How did they record Internet use?

3. Read the Results section:

1. How much time, on average, did their participants spend on non-class-related purposes during class?
2. How did the authors measure academic use versus non-academic use of the Internet?
4. Read the Discussion section: What are the key points in the discussion section?
5. Read the Abstract: How well did the abstract summarize the entire article?
6. Review the references. Pick a reference and try to find it using only its title via (a) a standard Google search, (b) Google Scholar, and (c) *PsycINFO*. What did you notice about these three search strategies?

Study Terms

Test yourself! Define and generate an example of each of these key terms.

- [abstract](#) (p. 22)
- [citations](#) (p. 23)
- [discussion](#) (p. 23)
- [generalization](#) (p. 22)
- [introduction](#) (p. 22)
- [literature review](#) (p. 24)
- [method](#) (p. 22)
- [parsimony](#) (p. 19)
- [prediction](#) (p. 31)
- [PsycINFO](#) (p. 28)
- [references](#) (p. 23)
- [research hypothesis](#) (p. 30)
- [results](#) (p. 22)
- [theory](#) (p. 18)
- [Web of Science](#) (p. 28)

Review Questions

Test yourself on this chapter's learning objectives. Can you answer each of these questions?

1. What are five different methods for generating ideas for research?
Provide an example of each.
2. What are the two major functions of a theory?
3. What information does the researcher communicate in each of the sections of a research article?
4. Describe differences in the way past research is found when using *PsycINFO*, *Web of Science*, or Google Scholar, including pros and cons of each.
5. What is a research hypothesis? How does it differ from a prediction?
What is the relationship between a theory and a hypothesis?

Deepen Your Understanding

Develop your mastery of these concepts by considering these application questions. Compare your responses with those from other people in your study group.

1. Think of a few common-sense sayings about behaviour (e.g., “Spare the rod, spoil the child”; “Like father, like son”; “Absence makes the heart grow fonder”). For each saying, develop a hypothesis that is suggested by the saying and a specific prediction that could follow from the hypothesis (Gardner, 1988).
2. Choose one of the hypotheses you formulated in Question 1 and find existing research studies that relate to that topic using the online database from your library. Page 35
3. Recall that theories serve two purposes: (1) to organize and explain past observations, and (2) to generate new knowledge by making predictions for future studies, guiding our attention of where to look next. Identify a consistent pattern of behaviour for either yourself or somebody close to you (e.g., you consistently get into an argument with your sister on Friday nights). Generate two possible theories that could explain these patterns (e.g., because you work long hours on Friday, you’re usually stressed and exhausted when you get home; *versus* because your sister has a chemistry quiz every Friday afternoon and she’s not doing well on them, she is very irritable on Fridays). How would you gather evidence to determine which theory might be correct? Do the two theories make different hypotheses about how to change the behaviour? If they do, they are considered competing theories. Is one more parsimonious than the other? Now consider each of your two theories, one at a time. Supposing your theory had been supported by empirical evidence, how would you change these behaviour patterns (e.g., to increase or decrease their occurrence)?

4. Use *PsycINFO* to find one journal article on a topic you're interested in (use [Appendix D](#) for tips). Identify all six elements of the research article, and compare what you can find in each section. Note where the research was conducted and who the researchers were. Use the Internet to find out if the researchers have websites. Where do they work? What other articles have these researchers published?

Ethical Research



©Westend61/Getty Images

This cat wants to research bullying by being super mean to you, but is this ethical?

Learning Objectives

Keep these learning objectives in mind as you read to help you identify the most critical information in this chapter.

By the end of this chapter, you should be able to:

1. LO1 Discuss the three core ethical principles for research with human participants, as outlined in the *Tri-Council Policy Statement*.
2. LO2 List and describe some of the potential risks and benefits of research.
3. LO3 Describe why informed consent is used despite the potential challenges to obtaining it.
4. LO4 Describe the purpose and process of debriefing research participants.
5. LO5 Describe the function of a Research Ethics Board (REB).
6. LO6 Contrast the categories of risk to participants: exempt, minimal risk, and greater than minimal risk.
7. LO7 Describe how the “Three Rs” are used to minimize harm to animals in research.
8. LO8 Discuss professional ethics issues, including scientific misconduct and transparency reform.

Page 37Ethical decision-making is crucial for all aspects of research. This includes planning, conducting, analyzing, publishing, and evaluating research. As members of the academic community, students and faculty alike must act in accordance with professional ethics. In this chapter, we will emphasize the ways that researchers strive to treat research participants with respect, including institutionalized guidelines and monitoring procedures. We will also explore some broader ethical issues, including fraud and plagiarism. A note of warning: Although some guidelines will seem straightforward, applying them in new research contexts can be very challenging.

Were Milgram's Obedience Experiments Ethical?

Let us begin by considering a familiar example. As you might recall, Stanley Milgram (1963, 1964, 1965) conducted a series of experiments to study obedience to an authority figure. He placed an ad in a local newspaper in Connecticut, offering to pay \$4.50 (approximately \$46 in 2018 Canadian currency) to men to participate in a “scientific study of memory and learning” being conducted at Yale University. The participants reported to Milgram’s laboratory, where they met a scientist dressed in a lab coat and another participant in the study, a middle-aged man named “Mr. Wallace.” Mr. Wallace was actually a confederate (i.e., an accomplice) of the experimenter, but the participants didn’t know this. The scientist explained that the study would examine the effects of punishment on learning. One person would be a “teacher” who would administer the punishment, and the other would be the “learner.” Mr. Wallace and the volunteer participant then drew slips of paper to determine who would be the teacher and who would be the learner. The drawing was rigged, however: Mr. Wallace was always the learner, and the volunteer was always the teacher.

The scientist attached electrodes to Mr. Wallace and placed the teacher in front of an impressive-looking shock machine. The shock machine had a series of levers and the individual was told that when these levers were pressed, they would deliver shocks to Mr. Wallace. The first lever was labelled 15 volts, the second 30 volts, the third 45 volts, and so on up to 450 volts. The levers were also labelled “Slight Shock,” “Moderate Shock,” and so on up to “Danger: Severe Shock,” followed by red Xs above 400 volts.

Mr. Wallace was instructed to learn a series of word pairs. Then he was given a test to see if he could identify which words went together. Every time Mr. Wallace made a mistake, the volunteer was to deliver a larger shock as punishment. The first mistake was supposed to be punished by a

15-volt shock, the second by a 30-volt shock, and so on. The learner, Mr. Wallace, never actually received any shocks, but the participants in the study didn't know that. Mr. Wallace made mistake after mistake. When the volunteer "shocked" him with about 120 volts, Mr. Wallace began screaming in pain and yelled that he wanted out. If the volunteer wanted to quit—and this definitely happened, with participants becoming visibly upset by Mr. Wallace's pain—the experimenter told the participant that he could quit but urged him to continue. These encouragements followed a set series of verbal prods that stressed the importance of continuing the experiment.

Page 38

Although the study purportedly was about memory and learning, Milgram was actually interested in whether participants would continue to obey the experimenter, administering ever higher levels of shock to the learner. What happened? Approximately 65 percent of the participants continued to deliver shocks all the way to the maximum possible: 450 volts. This study (and Milgram's many extended replications) received a great deal of publicity, and the results challenged many of our beliefs about our ability to resist authority. The results have implications for understanding obedience in real-life situations, such as the Holocaust in Nazi Germany and the Jonestown mass suicide (see A. G. Miller, 1986). Moreover, recent replications of these studies suggest that many people in contemporary society continue to be vulnerable to a dangerous obedience to authorities (Burger, 2009; Doliński et al., 2017).

But what about the ethics of the Milgram studies? Researchers debated whether these studies were ethical immediately after they were published (Baumrind, 1964; Kaufmann, 1967), and Milgram's work has shaped the common practices we use in psychology to protect our participants (Korn, 1997).

Ethical Research in Canada

The Tri-Council and Its Policy Statement

In Canada, researchers and institutions adhere to the *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans* (Canadian Institutes, 2010). Tri-Council simply refers to the three federally funded research grant agencies: the Canadian Institutes of Health Research (CIHR), the Social Sciences and Humanities Research Council of Canada (SSHRC), and the Natural Sciences and Engineering Research Council of Canada (NSERC). In 1998 the Tri-Council published the first [Tri-Council Policy Statement \(TCPS\)](#), which became the first standard Canadian ethics code to guide all research involving humans; this replaced all prior guidelines. In 2010, the Tri-Council published the first major revision of the TCPS, known as TCPS2, and these guidelines will continue to develop. All institutions whose researchers receive funding from the Tri-Council must have a *Research Ethics Board* (REB) that reviews each study to ensure it adheres to the TCPS2 ethical guidelines. We will further discuss REBs and their relationship to the Tri-Council [later](#) in this chapter.



TRY IT OUT!

Many colleges and universities require students and faculty to ensure they are up-to-date on current policies by completing the *TCPS2 Course on Research Ethics*. This is an online tutorial freely available at <http://tcps2core.ca>.

Historical, Legal, and International Context

The 1949 Nuremberg Code, which was developed in response to horrific human experimentation during World War II, was a catalyst for international debate on how to respect human dignity in medical and

behavioural research (Interagency Advisory Panel on Research Ethics, 2009). It emphasized the importance of *informed consent* (see [below](#)), and paved the way for updated international codes of ethics, including the World Medical Association's Helsinki Declaration in 1964, which had a massive impact on the conduct of medical research (McNeill, 1993). Countries and professional scientific societies also began codifying ethical practices for all research involving humans. For example, the core principles of Canada's TCPS2 are consistent with those outlined in an earlier American document (National Commission for the Protection of Human Subjects of Biomedical and Behavioural Research, 1979). This American report defined the principles that have guided more detailed regulations, and informed the American Psychological Association's own Ethics Code (American Psychological Association, 2010a; discussed [later](#) in this chapter).Page 39

The TCPS2 also reminds researchers to consult and follow the laws of the jurisdictions in which the research is conducted. In addition to adhering to the TCPS2, researchers must comply with the *Canadian Charter of Rights and Freedoms*, Canadian privacy of information laws, and relevant provincial laws. Varying legal contexts is one of the reasons why the TCPS2 is considered a set of guidelines rather than a set of rules. These guidelines are meant to provide guidance to researchers on how to behave, while also providing the flexibility needed to be useful across a wide variety of situations. Adopting guidelines rather than specifying rules acknowledges the reality of research as an innovative and constantly evolving enterprise. We can't have specific rules because we cannot predict every single specific situation posed by all future research studies. But we can develop guidelines to help us evaluate these novel future studies and situations as they arise.



Think about It!

Consider the example of a researcher who wants to collect data using Snapchat. Snapchat was released in 2011, after the TCPS2 was published, and so the TCPS2 doesn't include any mention of it. In fact, those who

wrote the TCPS2 probably didn't even conceive of this kind of technology, let alone set out rules governing research using this app. How should a researcher proceed in this instance? Does the TCPS2 apply, or help in any way, in this case?



LO1 Core Principles Guiding Research with Human Participants

The aim of research ethics codes around the world, including the TCPS2, is to ensure that research is conducted in a way that respects the dignity and inherent worth of all human beings. Three basic ethical principles express the value of ensuring human dignity and are specified in the TCPS2: (1) respect for persons, (2) concern for welfare, and (3) justice. The TCPS2 specifies the following:

- To show *respect for persons*, researchers must respect the autonomy of research participants, and protect those who have “developing, impaired, or diminished autonomy.” Respecting autonomy means enabling people to choose participation freely and without interference.
- To show *concern for welfare*, researchers must attempt to minimize risks associated with participating in research, while maximizing the benefits of that research to individual participants and to society.

When coupled with respect for persons, participants must be free to choose whether the balance of risk and benefits is acceptable to them.

- To show *justice*, researchers must treat people fairly and equitably by distributing the benefits and burdens of participating in research. Demonstrating justice includes recruitment methods that offer participation to people from a diverse range of social groups, and excluding groups only when scientifically justifiable.

These three principles provide ethical direction for all researchers who rely on human participants to help them make scientific discoveries. It is essential to continually consider the participant's perspective when applying these principles to research.

Designing Research to Uphold the Core Principles

Much of the TCPS2 is devoted to offering advice on how researchers and institutional Research Ethics Boards can implement the core principles effectively. Nonetheless, translating the core principles into practice can be challenging, with interpretation the source of ongoing discussion and debate. This section discusses some of the common research practices that promote each core principle, as well as related debates.



LO2 Promote Concern for Welfare by Minimizing Risks and Maximizing Benefits

The principle of [concern for welfare](#) refers to the need to maximize benefits and minimize any possible harmful effects of research participation. Think about the last time you made a big decision: Did you consider the relative risks (or costs) and benefits to yourself and others? In decisions about the ethics of research, we are required to calculate potential risks and benefits that are likely to result. This is called a [risk-benefit analysis](#).

Benefits to Participants and Society

Participants may experience several benefits from participating in research, including education about the scientific process, acquisition of a new skill, or treatment for a psychological or medical problem. They may also receive material benefits such as a monetary payment, a gift, the possibility of winning a prize, or points toward a course grade. Other, less tangible benefits may include satisfaction from contributing to a scientific investigation that could yield benefits for society. The knowledge gained through the research might also improve future educational practices, psychotherapy, or social policy.

Along with considering the potential benefits to society, it can be important to think about the cost of *not* conducting the study (cf., Christensen, 1988). This is especially true if the proposed research is the only way to collect potentially valuable data. For example, studying people's experiences of traumatic events may upset some participants, yet failure to study this topic can lead to misguided treatments and care (Newman, Risch, & Kassam-Adams, 2006). Importantly, although benefits and costs to society at large have a place in risk-benefit analysis, they are typically considered secondary to considering the ethical treatment of participants, the topic we turn to next.

Risk of Physical Harm

When it comes to risks to participants, perhaps the most obvious or salient is the potential for physical harm to participants. Researchers have affected participants physically in the pursuit of science across many different contexts. For example, researchers have administered alcohol to investigate the effects of intoxication on decision-making (Assaad et al., 2006), and have deprived people of sleep during different sleep phases to investigate effects on attention (Zerouali, Jemel, & Godbout, 2010). The risks of such procedures require that great care be taken to make them ethically acceptable. Moreover, there would need to be clear benefits of the research that outweigh the potential risks.

Risk of Psychological Stress

Aside from physical harm, participants may also experience psychological stress during research. Let's return to considering Milgram's research. It is not difficult to imagine feeling very stressed if you believe you are delivering intense shocks to an obviously unwilling learner. Footage exists of these studies by Milgram, and it shows participants protesting, sweating, and even nervously laughing while

delivering the shocks. When pursuing *concern for welfare*, we ask whether subjecting people to such a stressful experiment is justified, and whether the experience might have any long-term consequences for the participants. For example, would participants who obeyed the experimenter feel continuing remorse or begin to see themselves as cruel, inhumane people? A possible defence of Milgram's study follows, but first let's consider some other potentially stressful research procedures that have been used more recently.

Page 41

In past studies, participants were told that they will deliver a speech in front of a critical audience, receive only a few minutes to prepare the speech, and then go on to deliver the speech (Kirschbaum, Pirke, & Hellhammer, 1993). This method has been used to increase our understanding of physiological responses to stress among people with chronic major depressive disorder (Chopra et al., 2009; Harkness, Stewart, & Wynne-Edwards, 2011). Another potentially stressful form of research involves giving participants bogus tests of personality or ability. Using *deception* (see below), the researchers provide false feedback stating that the participant has an unfavourable personality trait or a low ability score. Not surprisingly, this kind of feedback leads participants to experience drops in self-esteem. But this procedure also allows researchers to examine the consequences of low self-esteem, such as how it affects well-being in the context of a romantic relationship (Cameron, Holmes, & Vorauer, 2009).

When stress is a possibility, even if just for a minority of participants, safeguards must be taken to help participants deal with the stress (Newman et al., 2006). Usually there is a *debriefing* session following the study that is designed in part to address potential stresses that may arise during the research (debriefing is discussed below).

Risk of Losing Privacy and Confidentiality

Another risk is the loss of expected privacy and confidentiality. Researchers must take care to protect the privacy of individuals, which includes "the right to control information about oneself" (Canadian Institutes, 2010). At a minimum, researchers should protect privacy by keeping all paper data locked in a secure place and encrypting all electronic data. Yet definitions of *privacy* and *confidentiality* are changing in the digital age (Richards & King, 2014), and those changing definitions will impact the way behavioural research is conducted. Ethical and legal policies are being redeveloped to accommodate this environment. For our purposes, it is important to be aware that using data for purposes other than what was agreed to during the *informed consent* process (see

below) may breach participants' privacy and confidentiality, adding risk to the participants in the form of loss of trust in researchers and the institutions they represent.

Confidentiality becomes particularly important when studying sensitive topics such as sexual behaviour, divorce, family violence, or drug abuse. For these topics, researchers may need to ask people very sensitive questions about their private lives. It is extremely important that individual responses to such questions be *confidential* (i.e., kept secret and used only for the purposes promised by the researcher). In many cases, the responses are completely *anonymous*—there is no way to connect any person's identity with any particular data. This happens, for example, when questionnaires are administered to groups of people, and no information is asked that could be used to identify an individual (such as name, student identification number, or phone number). In other cases, the identity of participants can be known, and therefore data are not completely anonymous. For example, it is usually impossible to guarantee complete anonymity in personal interviews or in online studies for which participants enter an e-mail address. The researcher must carefully design ways of coding data, storing data, and explaining the procedures to participants, to protect the confidentiality of responses and to ensure anonymity, when possible.

In some research, there is a real need to be able to identify individual participants. This occurs when people are studied on multiple occasions over time, or when specific personal feedback, such as an accurate test score, must be given to individual participants. In such cases, the researcher should create a code to identify the individuals, but should separate this code from the actual data. Thus, if questionnaires or data files were seen by anyone, the data could not be linked to specific people. In these cases where codes are used or participants can otherwise be identified, anonymity cannot be guaranteed, and researchers must take extra precautions to safeguard the confidentiality of all data.

Privacy laws in the United States can affect research conducted using the Internet in Canada. In particular, the *Patriot Act* in the United States allows the government to access records of Internet service providers. Therefore, online studies that are hosted by servers located in the United States risk the privacy and confidentiality of Canadian participants. For this reason, REBs may require that online studies are conducted using companies whose servers are located in Canada or outside of the United States. A similar issue occurs for research conducted over text messages or e-mail with servers hosted in the United States. Researchers who conduct studies online must develop or seek services that

provide safeguards such as encryption to protect participants' data from interception by unauthorized parties.

Another privacy issue concerns concealed observation of behaviour (see [Chapter 6](#)). In some studies, researchers make observations of behaviour in public places. Observing people in shopping malls (e.g., Ozdemir, 2008) or on sidewalks (e.g., Costa, 2010) does not seem to present any major ethical problems, as there is no reasonable expectation of privacy in these public places. However, what if a researcher wishes to observe behaviour online, in more private settings, or in ways that may violate individuals' privacy (see Wilson & Donnerstein, 1976)? For example, would it be ethical to rummage through people's trash or watch people in public washrooms?

In one famous study, Middlemist and colleagues (1976) measured how long it took men to start to urinate, and how long they urinated, in the washrooms at a college. The purpose of the research was to study the effect of personal space on anxiety, with urination times and delays serving as the measure of anxiety. The students were observed while alone or with an experimenter's confederate, who stood at the next stall or a more distant stall in the washroom. The presence and closeness of the confederate did have the effect of delaying urination and shortening the duration of urination. In many ways, this is an interesting study. It examines, for example, a situation that people experience on a regular basis. However, one can certainly question whether the invasion of privacy was justified, given the potential benefits (Koocher, 1977). To explore these concerns, the researchers used pilot studies, role-playing, and discussions with potential participants and eventually concluded that any ethical problems with the study were minimal (Middlemist, Knowles, & Matter, 1977).



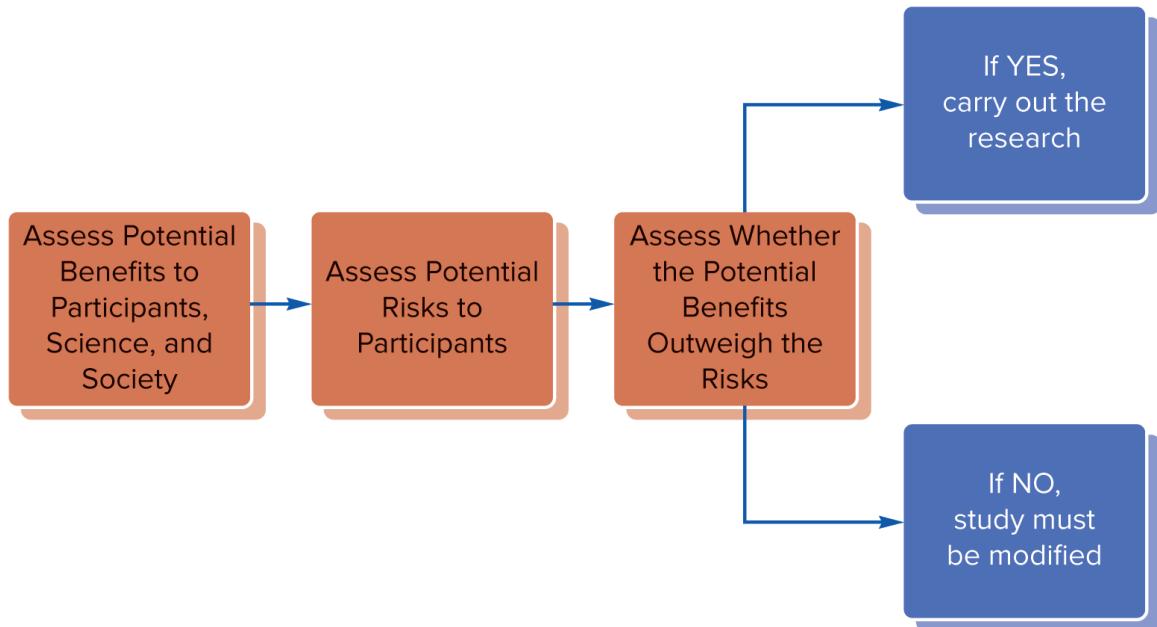
Think about It!

Do you agree that the ethical problems of this study by Middlemist and colleagues (1977) are minimal? Do the benefits outweigh the costs? What do you think of their approach to evaluating the potential risks?

To uphold the TCPS2 principle of *concern for welfare* means that researchers seek to minimize risks and maximize benefits to participants. [Figure 3.1](#) depicts a decision tree to help think through these complex issues, considering also the impact on society. Once a study has sufficiently demonstrated concern for

welfare, the next principle comes into action: Potential participants must be given the opportunity to assess their comfort with the risks and benefits involved. Page 43

Figure 3.1 Analysis of risks and benefits



LO3 Promote Respect for Persons through Informed Consent

The TCPS2 principle of [respect for persons](#) states that participants are treated as autonomous, capable of making deliberate decisions about whether they wish to participate in research. This is most often applied in the form of [informed consent](#). Potential participants in a research project should be provided with all information that might influence their decision to participate or not. For example, research participants should be informed about the purposes of the study, the risks and benefits of participation, and their rights to refuse or terminate participation in the study at any time without penalty. With this information, they can then freely consent to participate in the research, or refuse to do so, and they can also stop participating at any time knowing they won't be penalized in any way.

Informed Consent Form

Potential participants are usually provided with some type of informed consent form that contains the information they need to make their decision. There are numerous examples of informed consent forms available on the Internet. Your university may have examples available through its research office. A checklist for creating an informed consent form is provided in [Figure 3.2](#). Note that the checklist addresses both content and format. The content typically will cover (1) the purpose of the research; (2) the procedures that will be used, including time involved; (3) the risks and benefits to the participant and society in general; (4) any compensation; (5) how confidentiality will be protected; (6) the assurance of voluntary participation and permission to withdraw without penalty; and (7) contact information for those who have questions about the research and its ethics. Researchers do not need to tell participants exactly what is being studied, but the consent form must include all information that could affect a participant's choice to participate.

Figure 3.2 Checklist for informed consent form

Check to make sure the informed consent form includes the following:

- _____ Statement that participants are being asked to participate in a research study
- _____ Explanation of the purposes of the research in clear language
- _____ Expected duration of the subject's participation
- _____ Description of the procedures
- _____ Description of any reasonably foreseeable risks or discomforts and safeguards to minimize the risks

- _____ Description of any benefits to the individual or to others that may reasonably be expected from the research
- _____ If applicable, a disclosure of appropriate alternative procedures or courses of treatment, if any, that might be advantageous to the individual
- _____ Description of the extent, if any, to which confidentiality of records identifying the individual will be maintained
- _____ If an incentive is offered, a description of the incentive and requirement to obtain it; also, a description of the impact of a decision to discontinue participation
- _____ Contact information for questions about the study (usually phone contacts for the researcher, faculty advisor, and the Research Ethics Board office)
- _____ Statement that participation is voluntary, refusal to participate will involve no penalty or loss of benefits to which the subject is otherwise entitled, and the subject may discontinue participation at any time without penalty or loss of benefits to which the individual is otherwise entitled
- _____ Form is printed in no smaller than 11-point type (no “fine print”)
- _____ Form is free of technical jargon and written at a level appropriate for Grade 6 to 8 students
- _____ Form is not written in the first person (statements such as “I understand . . .” are discouraged)
- _____ Description of how data will be used and stored, including any foreseeable future uses (e.g., contribution to larger database)
- Other information may be needed for research with high-risk or medical procedures.

A sample informed consent form can be downloaded from  **connect**.

Page 44 There have been cases in which a consent form is so technical or loaded with legal terminology that it is very unlikely that participants fully realized what they were signing. In general, consent forms should be written in simple and straightforward language that avoids jargon and technical terminology. This generally means a Grade 6 to 8 reading level (word processing software may provide readability metrics, such as grade level, as part of a grammar check feature). To make the form easier to understand, it should not be written in the

first person. Instead, information should be provided as if the researcher were simply having a conversation with the participant. For this reason, sentences like, “Participation in this study is voluntary. You may decline to participate without penalty” are preferable to, “I understand that participation in this study is voluntary. I may decline to participate without penalty.” The first statement is providing information to the participant in a straightforward way using the second person (“you”), whereas the second statement has a legalistic tone that may be more difficult to understand. Other measures may need to be taken to help ensure that potential participants understand the form. For example, if participants are not fluent in the language the form is written in, there should be a translated version of the form. Researching special populations means taking additional considerations into account when designing the consent form. For example, researchers have developed careful methods to help drug users better understand the research to which they are consenting (e.g., HIV vaccine trials; Fisher, 2010).

There are research procedures in which informed consent is not necessary or even possible. If you choose to observe the number of same-gender and mixed-gender study groups in your library, you probably don’t need to announce your presence or obtain anyone’s permission. This is because you are not influencing or manipulating the people you are observing in any way, and everyone in the library understands that they may be observed by others in this space. However, when planning research without informed consent, it is important to make sure that you have good reasons for that decision. If informed consent is considered impossible to achieve the goal of science, the researcher’s responsibilities to protect the participants are increased.

Autonomy Issues

Informed consent seems simple enough, but there are important issues to consider. The first concerns lack of autonomy. What happens when the participants may lack the ability to make a free and informed decision to voluntarily participate? Special populations such as minors, patients in psychiatric hospitals, adults with cognitive impairments, or other vulnerable populations require special precautions. When minors are asked to participate, for example, a written consent form signed by a parent or guardian is generally required in addition to agreement by the minor. This agreement by a minor is formally called *assent*. The Division of Developmental Psychology of the American Psychological Association and the Society for Research in Child

Development have established their own guidelines for ethical research with children, which can be referenced for further details.

Coercion is another threat to autonomy. Any procedure that limits an individual's freedom to consent is potentially coercive. For example, a supervisor who asks employees to fill out a survey during a staff meeting is applying considerable pressure on potential participants. The employees may believe that the supervisor will somehow punish them if they do not participate. They also risk embarrassment if they refuse to participate in front of co-workers. Sometimes the compensation can be so great as to become coercive. For example, a prisoner may believe that increased privileges or even a favourable parole decision may result from participating. Or a person with extremely low income could be offered a large sum of money to participate in a study, and may feel pressure to participate out of financial need, even if she is uncomfortable with the risks involved. Researchers must consider these issues and make sure that autonomy is preserved.

Massive amounts of data are being created and stored as a result of our increased reliance on technology, such as smartphones, fitness trackers, smart devices, and the Internet as a whole. Many groups are interested in analyzing these data, including governments, businesses, and researchers. This *secondary use of data* can threaten people's ability to decide whether to participate in research. When you click "I accept the terms and conditions" when you sign up for an online service, is that equivalent to informed consent for any use of your data? Consider these examples of how you might be contributing archival data that could be used for secondary analyses: text and search histories, file downloads and uploads, music and e-book purchases, Netflix watching habits, e-mail, Facebook friends and followers, location information from your smartphone, and application usage patterns. Human behaviour is being captured constantly, raising deep and complex questions about the ethical use of these data.

As an example, consider a controversial experiment that involved collaboration between researchers working at Facebook (a profit-based company) and a Cornell university researcher. In 2012, unbeknownst to users, the Facebook researcher (a social psychologist) manipulated the emotional content of around 690,000 users' news feeds. Anonymous data were provided to researchers at Cornell, and the whole team analyzed the data and published a peer-reviewed research paper in a high-profile journal (Kramer, Guillory, & Hancock, 2014). Controversy erupted (Albergotti & Dwoskin, 2014; Felten, 2014), causing the journal editor to issue an "Editorial Expression of Concern" (Verma, 2014).



Think about It!

Why was this study by Facebook and Cornell researchers controversial (Kramer et al., 2014)? Do you consider this study a violation of one or more of the core principles for research ethics? If so, which one(s) and why? If not, why not?

Page 46

Withholding Information

All ethics codes since the Nuremberg Code have stressed the importance of informed consent as a fundamental part of ethical research. However, it may have already occurred to you that providing all of the relevant information about a study might be problematic. Providing too much information could potentially invalidate the results of the study if knowing all the details would change a person's behaviour so that it no longer resembles how that person would normally behave. Researchers will often withhold information about their hypothesis for a study, or the particular experimental condition in which the person is participating (see Sieber, 1992). It is generally acceptable to withhold this information when the information does not seem likely affect a person's decision to participate and when the information will be provided later, usually during a debriefing after the study is completed (see below).

Deception

Deception can sometimes be used as a tool by researchers when they actively misrepresent information to participants. Milgram's experiments illustrate two types of deception: (1) misleading participants about a study's goals, and (2) staging a situation. First, there was deception about the purpose of the study. Participants in Milgram's experiments agreed to take part in a study on memory and learning, but they actually took part in a study on obedience. What participant could have imagined that a memory and learning experiment would involve asking them to deliver painful electric shocks to another person? Because participants were deceived, they didn't know what to expect. Milgram's experiments also illustrate a second type of deception: Participants took part in a series of events that were orchestrated for the purposes of the study, like being an unknowing participant in a staged play. A confederate of the experimenter played

the part of another participant, and Milgram created a reality for the participant in order to observe obedience to authority.

Researchers typically engage in the first type of deception by creating a false purpose for the study, known as a cover story, to make the experiment seem both plausible and involving (Gross & Fleming, 1982; Hertwig & Ortmann, 2008). For example, a researcher might tell participants that they are reading boring newspaper articles for a study on readability, when the true purpose is to examine mind-wandering. Note that deception is not limited to laboratory research. When observing behaviours out in the real world, procedures in which observers conceal their purposes, presence, or identity are also deceptive (see [Chapter 6](#)).

Why do researchers use deception? Sometimes researchers are concerned that if participants know the true purpose of the study, they will not behave in a realistic fashion and this will affect the outcome of the study. In these cases, they wish to disguise the true purpose of the study during the informed consent procedure. Confirming these concerns, there is indeed research demonstrating that providing informed consent can bias participants' responses, at least in some research areas (Dill, Gilden, Hill, & Hanslka, 1982; Gardner, 1978). In Milgram's case, knowledge that obedience was being studied would likely have altered whether participants behaved in an obedient manner or not. Milgram's study was conducted before informed consent was routine, yet Burger's (2009) partial replication of this study suggests that fully informed consent would have changed the results. In this replication, participants who had previously learned about Milgram's work provided data that was so different from the others that it had to be excluded from analysis.

Page 47

It is also possible that the informed consent procedure can bias who ends up participating in a study, altering the characteristics of the sample. In Milgram's experiment, if participants had prior knowledge that they would be asked to give severe shocks to another person, some might have declined to participate. In fact, you could imagine that only a very specific kind of person might agree to shock other people in the name of science. By gathering data only from a particular sort of person, this would limit our ability to generalize the results more broadly. In this case, we would only be learning about the kinds of people who would agree to participate in such a study. If this were true, a critic might point out that the obedient behaviour seen in the Milgram experiment occurred simply because the people who agreed to participate were sadists in the first place! Researchers use deception when they are concerned that fully informed consent may affect who agrees to participate and/or how they behave once the study begins.

☆ Student Spotlight: Disclosure and Selection Bias ☆

An honours thesis student at the University of Calgary, Erin L. Moss, examined whether the information provided to potential participants might influence who chooses to participate in the context of eating research. Working with a supervisor, Dr. Kristin M. von Ranson, Erin designed an experiment that recruited women to participate in a study advertised with either one of two titles: “Disordered Eating in Young Women” or “Consumer Preferences.” They then examined whether they observed different results depending upon whether the participants who chose to participate knew in advance that the study was about disordered eating (Moss & von Ranson, 2006).



TRY IT OUT!

What do you think? Does knowing that a study is about eating disorders change the kinds of results you observe in a study? Track down the original article using the reference to find out for yourself!

Practical Difficulties of Deception

Although some researchers do use elaborate deception like in Milgram’s original study, these have become far less common since the 1980s (Hertwig & Ortmann, 2008; Sieber, Iannuzzo, & Rodriguez, 1995). There could be four potential reasons for this decrease in the use of elaborate deception. First, researchers have become more and more interested in cognitive variables and so use the methods found in memory research and cognitive psychology. Second, the general level of awareness of the ethical issues described in this chapter has led researchers to come up with alternatives to using deception (described further below). Third, ethics committees at universities and colleges (i.e., REBs) now review proposed research more carefully, so elaborate deception is likely to be approved only when the research is important and there are no alternatives available. Fourth, such elaborate setups are very difficult to achieve successfully, sometimes because participants are suspicious of the possibility of deception after learning about it in their courses (Hertwig & Ortmann, 2008). However, the effect of suspicion on a study’s outcome can also depend on what is being measured. Some physiological measures have shown no difference between people who are suspicious about deception and people who are not (Linden, Talbot Ellis, &

Millman, 2010). Nonetheless, elaborate deception is sometimes ruled out as impractical.



LO4 The Importance of Debriefing

A debriefing occurs after the completion of the study. This is an opportunity for the researcher to deal with issues of deception, withholding information, and potential harmful effects of participation. It is also a chance for the researcher to further educate participants about the nature and purpose of the research. Researchers are required to debrief participants when using deception or when they only partially disclose the purpose of the study (Canadian Institutes, 2010).Page 48

If participants were deceived in any way, the researcher needs to explain why the deception was necessary. If the study altered a participant's physical or psychological state in some way (e.g., a study that produced stress in order to study it), the researcher must make sure that the participant has returned to his or her original state and is comfortable with having participated. If the participant wants to receive additional information or to speak with someone else about the study, the researcher should be prepared to provide access to these resources. The participants should leave the experiment without any ill feelings toward the field of psychology, as research suggests they may become distrustful of researchers after experiencing deception (Blatchley & O'Brien, 2007). After a thorough debriefing, participants may even leave with some new insight into their own behaviour or personality.

The debriefing also provides an opportunity for the researcher to explain the purpose of the study, what kinds of results are expected, and any foreseeable

practical implications of the work. In some cases, researchers may contact participants later to inform them of the actual results of the study. Thus, debriefing serves both an educational and ethical function.

Is debriefing sufficient to remove any negative effects when stress and elaborate deception are involved? Little research has investigated the impact of debriefing in general (Sharpe & Faye, 2009), but what evidence does exist concludes that the debriefing effectively deals with deception and other ethical issues (Oczak, 2007; Smith, 1983; Smith & Richardson, 1983). In the case of Milgram's research, a thorough debriefing session was provided. Participants who were obedient were told that their behaviour was normal given the strong situational pressures; they had acted no differently from most other participants. Participants were assured that no shock was actually delivered, and there was a friendly reconciliation with the confederate. Milgram later mailed participants a report of his research findings along with a survey about their reactions to the experiment. Most participants (84 percent) reported they were glad that they had participated, and 74 percent said they had benefited from the experience. Only 1 percent said they were sorry they had participated. When a psychiatrist interviewed participants a year later, no ill effects of participation could be detected. Some who have evaluated Milgram's methods concluded that the debriefing helped his participants (Ring, Wallston, & Corey, 1970).

Alternatives to Deception

Given the risks of deception, it is important to evaluate whether it is necessary for a particular study. The Windsor Deception Checklist was developed to help researchers and REB members decide whether the use of deception for a particular study raises ethical concerns and should be reconsidered (Pascual-Leone, Singh, & Scoboria, 2010). If deception is considered sufficiently problematic for a study, a researcher may consider alternatives such as role-playing, simulation studies (a variation on role-playing), and “honest” experiments.

Role-Playing

Instead of using deception, experimenters might ask participants to role-play instead. In one role-playing procedure, for example, the experimenter describes a situation to participants and then asks them how they imagine they would respond to the situation (Kelman, 1967). Sometimes participants are asked to predict how real participants in such a situation would behave. Role-playing is

not generally considered to be a satisfactory alternative to deception (Freedman, 1969; A. G. Miller, 1972). One problem is that simply reading a description of a situation does not adequately convey what it would be like to really be in that situation. Also, because the experimenter provides a complete description of the situation, the hypothesis may become obvious to participants. When people can figure out the hypothesis of study, they may not behave as they normally would, but instead try to behave in a way that is consistent with the hypothesis (see *demand characteristics*, [Chapter 9](#)).Page 49

The most serious defect of role-playing is that no matter what results are obtained, critics can always claim that the results would have been different if the participants had been in a real situation. This criticism is based on research showing that people aren't always able to accurately predict their own behaviour or the behaviour of others (Aknin, Norton, & Dunn, 2009; Nisbett & Wilson, 1977). In fact, Milgram asked a group of psychiatrists to predict the results of his study and found that even these experts could not accurately anticipate what would happen. Nonetheless, role-playing can yield some interesting results when there are no reasonable alternatives. For example, Canadians who vividly imagined what it would be like to live with many social restrictions showed increased empathy for visible minority group members (Hodson, Choma, & Costello, 2009).

Simulation Studies

A different type of role-playing involves simulating a real-world situation. Simulations can be used to examine conflict between individuals in a competition, driving behaviour using driving simulators, or jury deliberations, for example. Such simulations can be highly involving and can effectively mimic many elements of a real-life experience. (See Mayhew et al., 2011, for a comparison of driving simulators with real driving.)

You may recall from your Introduction to Psychology class that one simulation study was so involving that it had to be terminated early: the Stanford Prison Study (Zimbardo, 1973; see also Haney & Zimbardo, 1998). In this study, student participants were paid \$15 per day (approximately \$112 in 2018 Canadian currency) and were randomly assigned to act as either guards or prisoners in a simulated prison, located in the basement of the psychology building at Stanford University. Participants became so deeply involved in their roles that the simulation was stopped after only six days (instead of two weeks) because of the cruel behaviour of the “guards” and the stressful reactions of the “prisoners.”

This was only a simulation—participants knew that they were not really prisoners or guards—yet they became so involved in their roles that the study produced extremely high levels of stress. The Stanford Prison Study is an unusual case, however. Most simulation studies do not raise the ethical issues seen in this particular study.

Honest Studies

Another alternative to deception is conducting an “honest” study, which is any research design that does not try to misinform or hide information from participants (Rubin, 1973). For example, studies of actual speed-dating events have been a very useful way to study romantic attraction (Finkel, Eastwick, & Matthews, 2007; Provost, Kormos, Kosakoski, & Quinsey, 2006). Participants can be recruited to engage in a speed-dating setting held on campus or at a local restaurant, where they complete numerous questionnaires and make choices that can lead to possible dates. This design enables the systematic examination of many factors related to dating without deceiving anyone about the purpose of the study. Another example of an honest design involves recruiting people who are seeking out information or a service. Students who wish to volunteer for a study skills improvement program could be assigned to either an in-class or an online version of the course. The researcher can then evaluate whether one version is superior to the other, all without using deception. Page 50

Another strategy that can avoid deception entirely involves naturally occurring events that present unique research opportunities. For example, Baum and colleagues (1983) studied the stressful effects of nuclear power plant disasters by asking people about their experiences living near a nuclear plant that had, or had not, accidentally released radioactive gases. The results of this study were actually replicated when researchers examined stress associated with the September 11, 2001, terrorist attacks (Schlenger et al., 2002). Science depends on the *replicability* of results (see [Chapter 14](#)), so this was an important demonstration using two very different, but both stressful, events in the real world. These studies illustrate how naturally occurring events can be valuable sources of important data.

Promote Justice by Involving People Equitably in Research

The principle of *justice*, as defined in the TCPS2, addresses issues of fairness in receiving the benefits of research as well as bearing the burdens of research risks. Any decisions to include or exclude certain people from a research study must be justified on scientific grounds. If age, ethnicity, gender, or other criteria are used to select participants, there must be a scientific rationale. The history of research includes too many examples of high-risk studies that were conducted on individuals selected because they were powerless and marginalized within society. One of the most horrific was the Tuskegee Syphilis Study in which 399 poor African Americans in Alabama were not given medical treatment in order to track the long-term effects of this disease (Reverby, 2000). This study took place from 1932 to 1972, ending when the details of the study were made public. The outrage over this study's targeted exploitation of a disadvantaged social group spurred scientists to overhaul ethical regulations in both medical and behavioural research. Yet examples continue to emerge. In Canada between 1942 and 1952, nutritionists and the federal government experimented with nutrients and food provided to malnourished members of Indigenous communities and residential schools—all without their consent (Mosby, 2013). Targeting a specific group of people to bear the risks of research is completely unacceptable.

The benefits of research must also be equitably shared across groups. It is unethical for a researcher to study a particular group of people to advance science if that particular group will not benefit from that research. Researchers from relatively wealthy nations have often investigated the health and psychology of people in developing nations. These research participants in developing nations should directly benefit from what is learned, but this has not always been the case (Glantz, Annas, Grodin, & Mariner, 2001). In Canada, the Tri-Council has made a clear statement that research with First Nations, Inuit, and Métis peoples must engage and benefit the involved communities. In addition to extending knowledge, whenever possible, research should be relevant to and “benefit the participating community (e.g., training, local hiring, recognition of contributors, return of results)” (Canadian Institutes, 2010, Article 9.13). Training and hiring members of a researched community are just some of the tangible and immediate ways that research can benefit participants and an entire community. A full chapter in the TCPS2 expands on how researchers can uphold all three core principles when working with Indigenous peoples.

Evaluating the Ethics of Research with Human Participants

When you are making decisions about how to treat research participants ethically, you can use the core principles of the TCPS2 as a compass. To put all three principles into practice, researchers need to consider many factors. Who are the participants? How will the results be shared with the participant community? Are there any risks of physical or psychological harm, or loss of confidentiality? What types of deception, if any, are used in the procedure? How will informed consent be obtained? What debriefing procedures are appropriate? Researchers also need to weigh the direct benefits of the research to the participants, the scientific importance of the research, and possibly the educational benefits to students who may be participating for a class requirement (see [Figure 3.1](#)).Page 51



Think about It!

These are not easy decisions. Consider a study in which a male confederate insults a male participant who grew up in either the northern or southern United States (Cohen, Nisbett, Bowdle, & Schwarz, 1996). The purpose was to investigate whether males in the South had developed a “culture of honour” that expects them to respond aggressively when insulted. Do you think that the potential benefits of the study to society and science outweigh the risks involved in the procedure?

For any study, if you ultimately decide that the costs outweigh the benefits, you must conclude that the study cannot be conducted in its current form. There may be alternative procedures that could be used to make it acceptable, reducing the risks or costs, or increasing the benefits. If the benefits outweigh the costs, you will likely decide that the research should be carried out. However, your calculation might differ from another person’s calculation, which is precisely why having Research Ethics Boards, discussed next, is such a good idea.



LO5 Monitoring Ethical Standards at Each Institution

Individual researchers working at Canadian universities often compete for research funding from one of the Tri-Council agencies. A researcher will typically apply to one of the three agencies, with this choice depending on the type of research being proposed. Once a researcher is awarded a research grant, the agency transfers the funds to the university, which in turn administers these funds for the researcher to use for that project. Individual researchers are ultimately responsible for complying with the ethical standards in the TCPS2. However, just as the Tri-Council relies on individual universities to administer research funds, it also involves universities in the process of ensuring that the TCPS2 is upheld in funded research. Each institution that receives any funding from any of the Tri-Council agencies must have a [Research Ethics Board \(REB\)](#), whose mandate is to review all research projects for compliance with ethical standards. Before starting any research project involving humans, researchers must apply for and receive ethical approval from their institution's REB. All research involving participants that is conducted by faculty, students, and staff associated with the institution is reviewed in

some way by the REB. This includes research that may be conducted at another location, such as a school, a community agency, a hospital, or online. Systematic review of research proposals reduces the chance that research that treats participants unethically will be conducted.

Consistent with regulations in other countries (e.g., U.S. Department of Health and Human Services) and professional societies (e.g., Canadian Psychological Association), the TCPS2 categorizes research according to the amount of risk involved. Researchers need to understand the level of risk to participants that their research entails. A brief description of each risk level and examples are provided below. However, because individual institutions' REBs will differ in how these categories are determined, and the details of each proposed study must be examined individually and in context, it is difficult to provide concrete rules for determining what kinds of research would fall in each category. Consult your institution's REB to learn precisely how it defines these categories.

Page 52



LO6 Exempt Research

Research in which there is absolutely no risk to participants is typically exempt from REB review. Researchers often are not permitted to decide by themselves that research is exempt from review. Instead, the institutional REB may create a procedure to allow a researcher to apply for this status. According to the TCPS2, exempt research does not require REB review when it (1) only uses publicly available information that is legally

accessible; (2) only involves observing people in public places without any intervention or interaction by the researcher, and no individuals can be identified when presenting the results; or (3) uses data that have already been collected and are completely anonymous. For example, archival research that uses anonymous data from Statistics Canada public use files (e.g., Stermac, Elgie, Dunlap, & Kelly, 2010) or published parliamentary archives (e.g., Suedfeld & Jhangiani, 2009) would be considered exempt. Also, *naturalistic observation* in public places when there is no threat to anonymity and no expectation of privacy would also be exempt. This type of research does not require informed consent. However, there is a continuing debate regarding exempt status in the context of big data.

Minimal Risk Research

According to the TCPS2, *minimal risk research* means that the risks of harm to participants are no greater than the risks one would normally encounter in daily life. The TCPS2 asks researchers to err towards a conservative interpretation of what constitutes “daily life,” in the interest of protecting participants. Minimal risk research still adheres to the core principles, but elaborate safeguards are not of great concern, and approval by the REB can often be delegated to a single member rather than considered by the whole committee. Individual REBs may vary in the specific criteria they use to determine minimal risk, but will likely consider the type of risk, the probability that it will occur, the amount of risk, and the vulnerability of the intended participants. For example, the minimal risk designation may apply to research using questionnaires or interviews with competent adults on non-sensitive topics (e.g., using questionnaires to measure the Big Five personality traits).



Think about It!

Throughout the course of our daily life, we can encounter all kinds of stressful and disturbing experiences. On any given day, we might learn that a beloved pet has passed away in an accident, for example. Does this mean that a research study in which participants are led to believe that their pet

has died would constitute minimum risk research? Why or why not? What other issues need to be considered?

Greater Than Minimal Risk Research

Any research procedure that places participants at *greater than minimal risk* is subject to thorough review by the full REB committee. In addition to informed consent, other safeguards may be required before approval is granted. Using questionnaires or interviews would be classified as greater than minimal risk if the topic was of a sensitive nature (e.g., illegal drug use), or if the intended participants were from a vulnerable population. In cases where it is ambiguous whether the research is minimal risk or greater than minimal risk, REBs will tend to categorize research as greater than minimal risk. This conservative approach reflects their mandate to uphold the core principles of respect for persons, concern for welfare, and justice.

Researchers planning to conduct an investigation that involves minimal risk or greater than minimal risk are always required to submit an application to the institution's REB. The specific application will vary by institution, but typically requires a description of risks and benefits, procedures for minimizing risk, procedures for recruiting participants, the exact wording of the informed consent form, how participants will be debriefed, and procedures for maintaining confidentiality and anonymity whenever possible. Even after a project is approved, there is often an ongoing review, with long-term projects reviewed at least once each year. If there are any changes in procedures or materials, researchers are required to obtain approval from the REB before implementing them.



LO7 Ethics and Animal Research

Although this chapter has been concerned with the ethics of research with humans, you are no doubt aware that psychologists sometimes conduct research with animals (Akins, Panicker, & Cunningham, 2004). Animals are used for various reasons. With animals, a researcher can carefully control the environmental conditions, study the same animals over a long period, and monitor their behaviour 24 hours a day if necessary. Animals are also used to test the effects of drugs and to study physiological and genetic mechanisms underlying behaviour in ways that would not be possible or ethical with humans. About 7 percent of the articles in the database *Psychological Abstracts* in 1979 described studies involving animals (Gallup & Suarez, 1985), and the amount of research done with animals has been steadily declining over time (Gauthier, 2004; Thomas & Blackman, 1992). Thus, animal research constitutes only a small proportion of psychology research conducted. In addition, within this small proportion of studies, over 95 percent of the animals used are rats, mice, and birds (see Gallup & Suarez, 1985).

Animal research, although controversial (Henry & Pulcino, 2009; Knight, Vrij, Bard, & Brandon, 2009; Plous, 1996a, 1996b), benefits humans and

continues to lead to discoveries that would not have been possible otherwise (Carroll & Overmier, 2001; N. E. Miller, 1985; see also www.apa.org/science/leadership/care/guidelines.aspx). Strict laws and ethical guidelines govern both research with animals and teaching procedures in which animals are used. These regulations stipulate the need for proper housing, feeding, cleanliness, and health care. Animal research must also avoid any cruelty in the form of unnecessary pain to the animal. Across all research with animals funded by the Tri-Council agencies—not just in psychology—the most common category of invasiveness is research that involves little or no discomfort or stress to the animals (Canadian Council on Animal Care, 2013).

The Canadian Council on Animal Care (CCAC) is an organization sponsored primarily by CIHR and NSERC. Its purpose is to “oversee the ethical use of animals in science in Canada” (Canadian Council on Animal Care, 2014). The CCAC is responsible for certifying institutions and specifying guidelines for ethical animal care for all research and teaching purposes across all disciplines (e.g., psychology, biology, medicine). In addition, institutions in which animal research is carried out must have an *Animal Care Committee* (ACC) composed of at least one scientist and/or teacher with experience in animal use, one institutional member who does not use animals, one experienced veterinarian, a community member, and others. The ACC is charged with reviewing animal research procedures and ensuring that all regulations are followed, including animal-care training by all researchers and technicians involved in animal research (see Canadian Council on Animal Care, 2006). This mandate includes promotion of the widely accepted *Three Rs* of “Good Animal Practice in Science” (Russell & Burch, 1959), which aim to minimize harm to animals:

1. *Replacement* involves replacing the use of animals with some alternative or avoiding the use of animals altogether (e.g., by using mathematical modelling).
2. *Reduction* involves minimizing the number of animals being used.
3. *Refinement* involves modifying procedures to minimize pain and distress.

Page 54For more information, see the CCAC's Three Rs website (<http://3rs.ccac.ca/en/>), which provides researchers with ideas and resources for following the Three Rs.

The Canadian Psychological Association (CPA) was an early leader among academic disciplines in creating ethical guidelines for the use of animals, now superseded by the CCAC guide. The American Psychological Association (APA) has also developed a detailed set of *Guidelines for Ethical Conduct in the Care and Use of Nonhuman Animals in Research* (American Psychological Association, 2012b). In addition to *replacing*, *reducing*, and *refining*, these APA guidelines state that psychologists using animals are expected to supervise all procedures and ensure that the comfort and health of animals is considered appropriately. Moreover, researchers must ensure that all personnel involved in the study (e.g., students, technicians, assistants) have received training in research methods, as well as how to properly care for the species involved. For more information about the ethics of animal research, please consult the resources on the CCAC website (www.ccac.ca) and your institution's Animal Care Committee.

☆ Student Spotlight: Animal Research ☆

Whenever working with animals, researchers strive to ensure that any harm or discomfort experienced by the animal is absolutely necessary. Although brain research using animals often employs the lesion method, in which brain tissue in a particular region is destroyed in order to uncover whether that region is necessary for a particular cognitive process, alternatives do exist. For example, it is possible to temporarily disrupt neural activity in a region rather than destroy it, which is a reversible process. This raises the question of whether the reversible, non-destructive approach is as effective as lesions in rendering a brain region inactive. Gavin Scott, an undergraduate working with Drs. Deborah Saucier (University of Ontario Institute of Technology) and Hugo Lehmann (Trent University), explored this question empirically. In order to do so, they applied both methods to the hippocampi of rats, and observed their response to a fear conditioning manipulation. Did they observe equivalent results across the two methods?

To find out, you can read their article in the *Journal of Behavioral and Brain Science*, (Scott, Saucier, & Lehmann, 2016).

Professional Ethics in Academic Life

Treating human participants and non-human animals ethically is just part of what it means to conduct research ethically. It is expected that psychologists behave ethically in all areas of professional activity, including publishing, teaching, clinical work, university administration, and communicating with the general public. As members of the academic community, students are also expected to behave ethically. In this section we will explore some further issues in ethical conduct within the professional realm.

Ethics Codes of the APA and CPA

Both the APA and the CPA have provided leadership in formulating *ethics codes*: ethical principles and standards for all aspects of a professional academic career in psychology. The APA Ethics Code provides an overview of how far professional ethics extends for all psychologists, beyond what we have considered in this chapter:

- Psychologists are committed to increasing scientific and professional knowledge of behavior and people's understanding of themselves and others and to the use of such knowledge to improve the condition of individuals, organizations, and society. Psychologists respect and protect civil and human rights and the central importance of freedom of inquiry and expression in research, teaching, and publication. They strive to help the public in developing informed judgments and choices concerning human behavior. In doing so, they perform many roles, such as researcher, educator, diagnostician, therapist, supervisor, consultant, administrator, social interventionist, and expert witness.

Both the APA and CPA Ethics Codes offer standards for ethical conduct by professional psychologists across all of the roles listed above (American Psychological Association, 2010a; Canadian Psychological Association, 2000). Koocher (2009) suggests that when prioritizing these various roles, the primary responsibility of psychologists is toward the most vulnerable members of society (rather than, for example, toward the government or a corporation). The APA Ethics Code is longer and more detailed than the CPA Ethics Code, yet the two complement each other. Canadian psychologists can benefit from relying on both

codes for ethical guidance, but the CPA Ethics Code is consistent with the TCPS2 and highlights the Canadian context of research, teaching, clinical work, and other activities.



LO8 Scientific Misconduct and Publication Ethics

Ethical decision-making in data collection, data analysis, and publication has recently taken centre stage in psychology, partly in response to some high-profile cases of scientific misconduct (Pashler & Wagenmakers, 2012). Fundamentally, we must be able to believe the reported results of research, otherwise the entire foundation of the scientific method as a means of knowledge is threatened. Scientific misconduct, in which researchers behave unethically to produce false evidence for a phenomenon, undermines the very nature of scientific investigation and the public's trust in science. Misconduct is possible because there are many points in the publication process where scientists have a great deal of freedom in how data is collected, analyzed, and reported. This flexibility allows for researchers to make decisions that falsely show evidence for a phenomenon. A key paper by Simmons, Nelson, and Simonsohn (2011) describes just how false evidence for an effect can be produced, and is recommended reading for all aspiring scientists. It has now become clear that we can no longer assume that scientific misconduct is extremely rare and of little concern (Stroebe, Postmes, & Spears, 2012), and all researchers are obligated to educate themselves on how to best conduct research in order to avoid acting unethically, including how to avoid the key problems described by Simmons and colleagues (2011).

Fabricating Data and Altering Collected Data Is Fraud

Two ways in which *[fraud](#)* is committed in science include (1) fabricating (i.e., making up) data and (2) collecting real data but altering the numbers to fit the hypothesis. The Ethics Codes of both the APA and CPA clearly prohibit these actions. Yet some researchers in psychology, biomedicine, chemistry, and other sciences have been caught doing both of these things (Stroebe et al., 2012). Three major cases in social psychology have received a great deal of media attention. In 2001, Dr. Karen Ruggiero (a former McGill University undergraduate) was caught after having fabricated data in many studies that were published in high-profile journals. Ruggiero subsequently resigned from her academic position and retracted her publications (Murray, 2002). More recently, Dirk Smeesters resigned after seven of his papers showed evidence of data fabrication or manufactured analyses (Report of the Smeesters Follow-Up Investigation Committee, 2014). An even more widespread act of fraud was discovered in November 2012, when it was found that Diederik Stapel committed fraud in 55 publications (from 1996 to 2011), including his own PhD dissertation and the dissertations of ten graduate students (Leveldt Committee, 2012). After having received many early career awards and great prestige for his work, his graduate students developed suspicions about the research being done in the lab, based on how perfectly Stapel's data always fit his predictions (for example, Stroebe et al., 2012). Three of his graduate students eventually reported Stapel to their department head and an in-depth investigation ensued. In the end, many affected papers were retracted, he was fired, he relinquished his PhD, and he underwent a criminal investigation that ended in a community service settlement (Bhattacharjee, 2013). Stapel, Ruggiero, Smeesters, and all others who have been caught committing scientific fraud have destroyed their own reputations and careers, and in some cases have hurt the reputations of their collaborators and students. The website [retractionwatch.com](#) is a good resource for exploring and keeping up to date with retracted papers across many different disciplines, including psychology, and includes a searchable [database](#) of retracted papers.

Page 56



Think about It!

When a researcher manufactures false evidence for a phenomenon, by either falsification or by taking advantage of the flexibility afforded to researchers when conducting research, many different parties are harmed. Try to list all the different people and institutions that would be harmed if a researcher engaged in scientific fraud, claiming evidence for something where none exists. What are the

consequences if the fraud is detected? If it goes undetected? If you think about it, the negative impact of fraud is felt widely with serious consequences for many.

How is fraud in science detected? It is possible that fraudulent results might be detected when other scientists fail to replicate the results of a study, or through the peer review process. However, data show that fraud is often detected by colleagues or students working with the researcher (as was the case with Ruggiero and Stapel; Stroebe et al., 2012). In addition, people are beginning to develop methods for detecting fraud in a paper by statistically analyzing the published results (Simonsohn, 2013). One of these methods was what triggered the investigation into Smeesters' work (Simonsohn, 2013). There are also methods being developed to detect whether a set of papers provide good evidence for a phenomenon, based on the observation that when a real effect exists, *p*-values just below .05 are rather unlikely (Simonsohn, Nelson, & Simmons, 2014; Simonsohn, Simmons, & Nelson, 2015).

Why do some researchers commit fraud? One possible reason is that scientists occasionally find themselves seeking or holding academic positions with extreme pressure to produce impressive results. This is not a sufficient explanation, of course, because many researchers maintain high ethical standards under such pressure. Another reason may be that researchers who feel a need to produce fraudulent data have an exaggerated fear of failure, plus a great need for success and admiration. If you wish to explore further the dynamics of fraud, a good starting point is Goodstein's (2010) *On Fact and Fraud: Cautionary Tales from the Front Lines of Science*.

Because scientific fraud is such a serious offence, allegations of fraud should not be made lightly. If you disagree with someone's results on philosophical, political, religious, or other grounds, it does not mean that they are fraudulent. Even if you cannot replicate the results of a study, there may be other reasons for why the original study reported an effect aside from deliberate fraud. However, the fact that fraud could be a possible explanation of results emphasizes to researchers the importance of keeping careful records and documentation of all procedures, analyses, and results to be able to prove that fraud did not take place.

Page 57

Ethical Data Analysis

As you may notice in [Chapters 12](#) and [13](#), the statistical analysis of data can be complex and include a level of judgment that may surprise you. Mistakes like

rounding errors can occur sometimes in published reports of analyses. Although it is difficult to know whether mistakes are accidental or intentional, data show that errors tend to occur in favour of the researcher's hypothesis, indicating that they are likely to be intentional (Bakker & Wicherts, 2011). Knowingly changing the numbers in your analysis is unethical and constitutes scientific fraud. To avoid unintentional errors, it is good practice to double-check your analyses as well as your final report to ensure accuracy. You can ask someone else whom you trust, and who is knowledgeable, to check them as well.

Improving Science through Publication Reform

Psychologists and other scientists are making progress as they seek to improve the trustworthiness of published research, but much hard work lies ahead (Miguel et al., 2014). Recent reform has emphasized improving transparency, allowing others to see exactly what was done at each stage of a research project (e.g., by supporting resources like the Open Science Framework: <https://osf.io/>). Three concrete practices that individual researchers can adopt if they want to help build a more honest and accurate science are as follows: (1) disclosure, (2) preregistration of studies, and (3) open data and materials (Miguel et al., 2014). Disclosing all information about a study (e.g., all measures used, reasons for excluding any participants) keeps a record of exactly how the research was conducted, facilitating replication attempts, and allows others to better evaluate your research. Preregistration involves posting a public statement of what your method will be as well as your data analysis plan, all before the data are collected. This prevents researchers from capitalizing on the freedom and flexibility we discussed earlier, reducing the possibility of making self-serving decisions regarding design and analysis that falsely create the appearance of evidence when none exists. Lastly, posting data files and study materials online will help other researchers evaluate a study as they will have all the needed details. This also allows others to replicate a paper's analyses, which may help catch fraud or unintentional errors (Simonsohn, 2013). Importantly, these practices are being promoted and sometimes mandated by editors of major journals (e.g., *Psychological Science*; Eich, 2014). These and other reforms will develop further as more researchers dedicate effort to the honest pursuit of truth via science, and it is every researcher's responsibilities to keep abreast of these developments and adopt them accordingly.

Plagiarism and the Integrity of Academic Communication

Plagiarism refers to presenting another person's work or ideas as your own, intentionally or even *unintentionally*, and is another form of serious scientific misconduct. All members of the academic community—students and professors alike—must give proper citation to all sources to signal where the ideas of others end and our own ideas begin. Plagiarism can take the form of submitting an entire paper written by someone else. It can also mean including a paragraph or even a sentence that is copied from another source without using quotation marks and a reference to the source of the quotation. Also, paraphrasing the words used by a source can become plagiarism if the source is not cited, as it often means copying a person's ideas or the structure of their argument. Learning how to fairly represent the ideas of others is a skill that is essential to the academic conversation—a conversation in which you engage every time you submit coursework.

Page 58

Although plagiarism is certainly not a new problem, access to Internet resources and the ease of copying material from the Internet may be increasing its prevalence. Szabo and Underwood (2004) found that more than 50 percent of a sample of British university students mistakenly believed that using Internet resources for academically dishonest activities, such as plagiarism, was acceptable. (To be clear, it is not.) Because of plagiarism concerns, many schools are turning to automated ways of detecting plagiarism (e.g., www.turnitin.com). Plagiarism is ethically wrong and can lead to many strong sanctions. These include academic sanctions such as a failing grade or expulsion from one's school. Because plagiarism is often a violation of copyright law, it can also be prosecuted as a criminal offence.

One reason why students engage in cheating may be poor writing ability (Williams, Nathanson, & Paulhus, 2010). Poor writing skill is *not* an acceptable excuse for plagiarism. Instead, seek assistance to improve your writing from your instructor, your institution's writing centre, and numerous websites, such as Purdue University's extensive Online Writing Lab (<http://owl.purdue.edu>). Another reason why students may plagiarize is the belief that citing sources weakens their papers—that they are not being sufficiently original. This is not at all the case. In fact, Harris (2002) notes that student papers are actually strengthened when sources are used and properly cited. Not knowing how plagiarism is defined is not an excuse to plagiarize, just as ignorance of a law does not protect you if you were to break that law. It is your responsibility to both learn what plagiarism is and how to avoid it. Take a look at [Figure 3.3](#) for a rough guide to avoiding plagiarism. Then practise identifying what constitutes plagiarism using the Test Yourself! in [Table 3.1](#).

Figure 3.3 Guide for avoiding plagiarism in writing

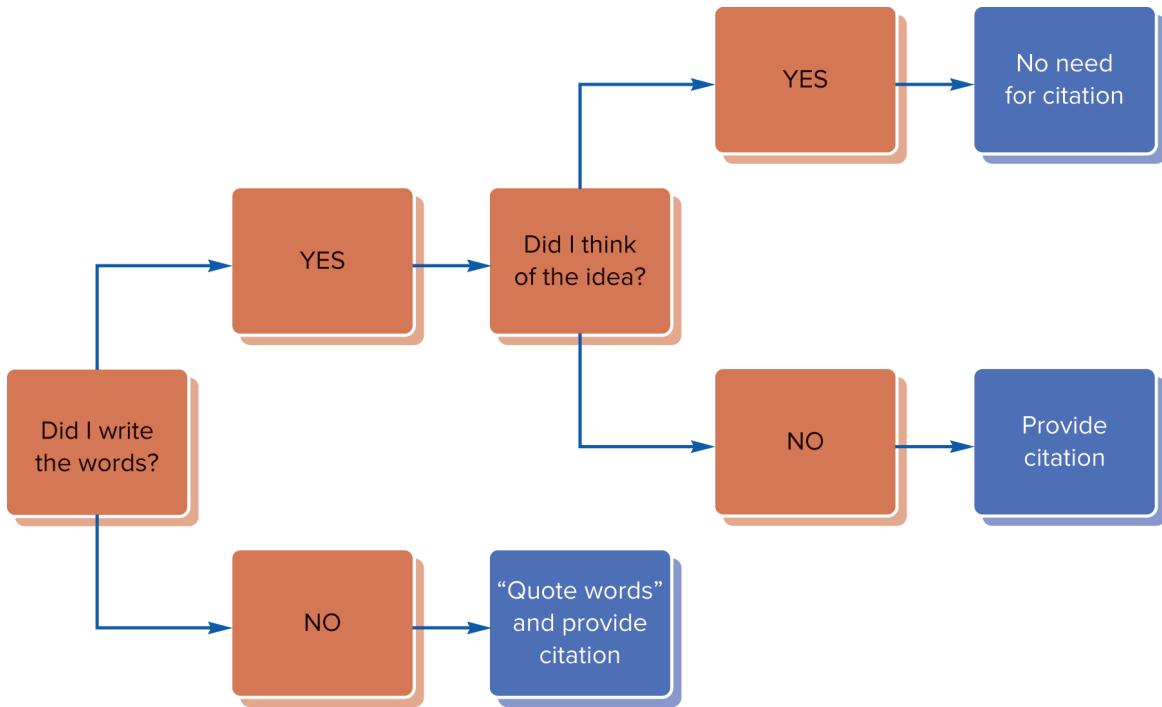


Table 3.1 TEST YOURSELF!: Plagiarism

Read the following examples and identify which counts as plagiarism and which does not. For each example you identify as plagiarism, specifically outline what changes you could make to avoid plagiarism. If you are unsure of any example, discuss it with your instructor, writing centre staff, or librarian.

Plagiarism	Not Plagiarism
1. Copying an entire essay and presenting it as original work.	

**Plagiarism Not
Plagiarism**

2. Compiling phrases, sentences, or paragraphs from a variety of sources to create a piece of work.
3. Taking words from another author without acknowledging that they are not your own.
4. Taking words from another author and noting that they are not your own (e.g., using quotation marks).
5. Paraphrasing ideas from another author without acknowledging that they are not your own.
6. Submitting work with incomplete source information (e.g., an incomplete References section).
7. Using material from other sources as if it were the results of your own research.

Page 59 Ethical guidelines and regulations are constantly developing. The TCPS2 and the APA and CPA Ethics Codes, as well as federal, provincial, and local regulations, are revised periodically. It is the responsibility of all researchers to be aware of the most current policies and procedures, and to ensure that they follow them. In the following chapters, we will discuss many different procedures for studying behaviour. As you read about these procedures, practise thinking about the ethical considerations that arise for each.



Illustrative Article: Replication of Milgram

Burger (2009) conducted a partial replication of the classic Stanley Milgram obedience studies.

First, acquire and read the article:

- Burger, J. M. (2009). Replicating Milgram: Would people still obey today? *American Psychologist*, 64, 1–11. doi:10.1037/a0010932

Then, after reading the article, consider the following:

1. Conduct an informal risk-benefit analysis. What are the risks and benefits inherent in this study as described?
2. Do you think that the study is ethically justifiable, given your analysis? Why or why not?
3. How did Burger screen participants in the study? What was the purpose of the screening procedure?
4. Burger paid participants \$50 for two 45-minute sessions. Could this be considered coercive? Why or why not?
5. Describe the risks to research participants in Burger's study.
6. Burger uses deception in this study. Is it acceptable? Do you believe that the debriefing session described in the report adequately addresses the issues of deception?

Study Terms

Test yourself! Define and generate an example of each of these key terms.

- [anonymous \(p. 41\)](#)
- [concern for welfare \(p. 40\)](#)
- [confederate \(p. 37\)](#)
- [confidential \(p. 41\)](#)
- [debriefing \(p. 47\)](#)
- [deception \(p. 46\)](#)
- [ethics codes \(p. 54\)](#)
- [exempt research \(p. 52\)](#)
- [fraud \(p. 55\)](#)
- [informed consent \(p. 43\)](#)
- [justice \(p. 50\)](#)
- [minimal risk research \(p. 52\)](#)
- [plagiarism \(p. 57\)](#)
- [Research Ethics Board \(REB\) \(p. 51\)](#)
- [respect for persons \(p. 43\)](#)
- [risk-benefit analysis \(p. 40\)](#)

- *secondary use of data* (p. 45)
- *Three Rs* (p. 53).
- *Tri-Council Policy Statement (TCPS)* (p. 38)

Review Questions

Test yourself on this chapter's learning objectives. Can you answer each of these questions?

1. Discuss the three major ethical principles in behavioural research and how they are related to risks, benefits, deception, debriefing, informed consent, and participant recruitment. How can researchers weigh the need to conduct research against the need for ethical procedures?
2. How does informed consent address respect for persons? What are the potential challenges involved in obtaining fully informed consent?
3. What are the purposes of debriefing participants after deception? What are some alternatives to deception?
4. What is a Research Ethics Board, and what does it do?
5. What are the differences among "exempt," "minimal risk," and "greater than minimal risk" research?
6. Summarize the ethical principles and procedures for research with animals.
7. What constitutes fraud? Why does it occur? What practices are being promoted to help avoid and detect it?
8. What activities do psychologists perform that require upholding professional ethical standards?

Deepen Your Understanding

Develop your mastery of these concepts by considering these application questions. Compare your responses with those from other people in your study group.

1. Consider the following experiment, similar to one that was conducted by Smith, Lingle, and Brock (1979). Each participant interacted for an hour with another person who was actually a confederate. After this interaction, both people agreed to return one week later for another session with each other. When the real participants returned, they were informed that the person they had met the week before had died. The researchers then measured reactions to the death of the person.
Page 61
 1. Discuss the ethical issues raised by the experiment.
 2. Would the experiment violate any of the three core ethical principles for research with human participants? In what ways?
 3. What alternative methods for studying this research question (reactions to death) might you suggest?

Now consider these same three questions regarding a study conducted with final-year medical students (Fraser et al., 2014). In a live, but simulated, situation, a 70-year-old woman had ingested a poison and was suffering from lack of consciousness. For half the participants, the patient unexpectedly died. The researchers measured participants' emotional responses and cognitive load after the event, and their competence in a similar simulated situation three months later. *How have your responses changed? Why?*

2. In a procedure described [earlier](#), participants are given false feedback that they have an unfavourable personality trait or a low ability level. What are the ethical issues raised by this procedure? Consider risks versus benefits. What if people are given false feedback that they

possess a very favourable personality trait or a very high ability level, instead of negative false feedback? Does that change your reaction to this method?

3. A social psychologist conducts a field experiment (i.e., an experiment that occurs outside of the laboratory) at a local bar that is popular with university students. Interested in observing flirting techniques, the investigator instructs male and female confederates to smile and make eye contact with customers at the bar for varying amounts of time (e.g., two seconds, five seconds, etc.) and varying numbers of times (e.g., once, twice, etc.). The investigator observes the responses of those receiving the gaze. What ethical considerations, if any, do you perceive in this field experiment? Is there any deception involved? Should these field experiment participants be debriefed? Write a list of arguments supporting the pro position and another list supporting the con position.
4. Assess the level of risk for the following research activities. Why did you choose that level of risk? Could you argue for a different level of risk?

No Risk	Minimal Risk	Greater Minimal Risk
Than		
Minimal Risk		Minimal Risk

1. Researchers conducted a study on a university campus examining the physical attractiveness level among peer groups by taking pictures of students on campus and then asking students at another university to rate the attractiveness levels of each student in the photos.

No Risk	Minimal Risk	Greater Than Minimal Risk
----------------	---------------------	----------------------------------

2. A group of researchers plan to measure differences in depth-perception accuracy with and without binocular cues. In one condition, participants could use both eyes, and in another condition, one eye was covered with an eye patch.

3. Researchers conducted an anonymous survey on attitudes toward gun control among shoppers at a local mall.

4. University students watched a ten-minute video recording of either a male or female newscaster presenting the same news content. While the video played, an eye movement recording device tracked the amount of time the students were viewing the video.

Chapter 4

Page 62

Research Design Fundamentals



©Ingram Publishing/SuperStock

This kitten is a bit confused about science right now, but don't worry, we're going to start with the basics.

Learning Objectives

Keep these learning objectives in mind as you read to help you identify the most critical information in this chapter.

By the end of this chapter, you should be able to:

1. [LO1](#) Compare and contrast non-experimental and experimental research methods.
2. [LO2](#) Define the term *operationalization* for a variable.
3. [LO3](#) Give examples of confounding variables and describe how to avoid them.
4. [LO4](#) Describe different possible relationships between variables: positive, negative, curvilinear, no relationship, and mediating.
5. [LO5](#) Distinguish between an independent variable and a dependent variable.
6. [LO6](#) Discuss how the three criteria for inferring causation are achieved (or not) in experimental and non-experimental methods.
7. [LO7](#) Describe why multiple methods are used for research, including the advantages and disadvantages of the basic designs.

Page 63 There are two major types of studies: experiments and non-experimental designs. Discussing them together will help to highlight important differences between the two types. We aim to offer enough detail here for an introduction to these types of studies, but leave some of the more complex issues for later chapters. First, it's important to discuss how researchers take variables of interest and incorporate them into studies. Then we unpack key concepts in non-experimental and experimental designs, respectively, and close by considering the value of using multiple types of studies.

Introduction to Basic Research Design

In [Chapter 2](#) we explored some ways researchers come up with ideas and begin to turn them into falsifiable hypotheses and predictions. In this section, we break down those steps even further by considering ideas as sets of variables, with these variables then incorporated into different research designs in order to ask research questions.

Variables

A research idea can be broken down into a set of variables. A *variable* is any event, situation, behaviour, or individual characteristic that can differ in some way (e.g., differ in size, degree, nature). In simplest terms, variables are things that vary (e.g., between people), things that can have more than one value. For example, a researcher interested in the effects of sleep on memory is interested in two variables: sleep and memory. Both the amount of sleep that people get and how good their memory is varies between individuals. Some people get more or less sleep, and some people have better or worse memory. Other examples of psychological variables include heart rate, intelligence, gender identity, reaction time, attractiveness, happiness, stress, sexual orientation, age, and personality traits. Each of these variables will have at least two specific *levels* or *values* (and in many cases, many more than two). For some variables, the values will have true numeric, or quantitative, properties or meaning. Suppose that memory is measured using a 50-question test on which scores can range from 0 correct to all 50 correct. These values have real numeric properties; the numbers mean something tangible. For other variables, the values may not be numeric, but instead simply identify different categories. An example of a categorical variable is nationality, the country in which people were born. These values are different, but they do not differ in amount or quantity. We will consider variable properties further in [Chapter 5](#).



LO1 Two Basic Research Designs

With variables of interest in mind, researchers must select a study design that will help them answer their questions about those variables. There are two general approaches to studying the relationships among variables: non-experimental methods and the experimental method. With a non-experimental method, relationships are studied by measuring or observing the variables of interest. This could include asking people to describe their behaviour, directly observing behaviour, recording physiological responses, or examining publicly available information (e.g., census data, participation on public websites). Once data on both variables are collected, the researcher uses statistics to determine whether there is a relationship between them. As values of one variable change, do values of the other change as well? For example, Gaudreau and colleagues (2014) measured how often undergraduate students used their laptops in class for unrelated tasks (e.g., visiting social network sites), and investigated whether this was related to academic grades. The two variables did indeed vary together: The more students used their computer in class (for unrelated things), the lower their grades.

Page 64
☆ Student Spotlight: Correlational Research ☆

While an undergraduate student at the University of Regina, Gregory P. Krätzig wondered whether people truly do learn better when material is presented in their preferred format. Is there such a thing as a visual learner, an auditory learner, or a tactile learner? In collaboration with Dr. Katherine D. Arbuthnott, they set out to investigate this question by asking people about their preferred learning style and then having them complete objective measures of learning using different modalities (e.g., visual, auditory). They then analyzed these data to see if those who reported preferring one type of learning (e.g., visual learners) actually performed better when learning with those materials (e.g., pictures). Their results now appear in the *Journal of Educational Psychology* (Krätzig & Arbuthnott, 2006).



TRY IT OUT!

In the Student Spotlight about correlational research, and subsequent ones, we haven't mentioned the full reference for the published article. It's good to practise tracking down an article when you don't have the actual reference, as references often don't appear when research is discussed in the popular press (e.g., newspaper and magazine articles). Without checking the list of references at the back of this book, can you find this article and discover the results of this study?

In contrast to non-experimental methods that measure all the variables of interest, the *experimental method* involves direct manipulation of one variable, control of several other variables (those not of interest), and then measurement of an outcome variable (hypothetically affected by the manipulated variable). The differences between non-experimental and experimental methods have important implications for the types of conclusions we can draw from the results.

☆ Student Spotlight: Experimental Research ☆

Chantal M. Boucher, working with Dr. Alan Scoboria at the University of Windsor, was curious about life transitions (e.g., moving to a new city to

attend university), and whether framing these transitions in different ways affects how we see them. They manipulated how people framed these transitions by asking people to either focus on the details of the transition (e.g., packing up your belongings) or the broader significance of the event for their life (e.g., starting on a new journey) while writing about it. After this manipulation, they measured several possible outcome variables (e.g., impact of the transition, relevance to the self). This research was subsequently published in the *Journal of Personality* (Boucher & Scoboria, 2015).

Page 65



LO2 Operationally Defining Variables: Turning Hypotheses into Predictions

Once a researcher decides on a method to study the variables of interest, these abstract variables must be translated into concrete and specific forms, for either observation or manipulation. Variables like “aggression” or “amount of reward” must all be defined in terms of the specific method that will be used to measure or manipulate it. Scientists call this the *operational definition* of a variable, or *operationalization*—a definition of the variable in terms of exactly what operations or techniques will be carried out in a particular study to represent that variable.

When considering options for operationalizations, sometimes it is helpful to think about variables as grouping into three general categories. A *situational variable* describes characteristics of a situation or environment to which the participant is exposed. Examples include the length of a passage being read, the number of people around you when writing a test, the credibility of a person trying to persuade you, or different instructions on what to focus on while recalling an event. Situational variables can be measured in any design or manipulated in experimental designs. A *response variable* refers to the responses or behaviours of participants, such as reaction time in response to a stimulus, performance on a cognitive task, and emotional reactions. Response variables are measured in either experimental or non-experimental designs. A *participant variable* describes a characteristic that individuals bring with them to a study, including cultural background, age, intelligence, and personality traits such as extraversion. Sometimes participant variables are related to other variables in non-experiments, and sometimes they are used to group participants for comparison on some response variable (see [Chapter 11](#)).

Variables must be operationalized so that they can be studied empirically. Recall from [Chapter 2](#) that hypotheses are more abstract than predictions. A prediction is a statement of the hypothesis that has been translated into the specific operationalizations of that particular study. Consider the hypothesis, “Hunger predicts aggression.” To test this hypothesis, researchers need to decide exactly what they mean by the terms “hunger” and “aggression” because there are many ways of operationalizing them. For example, aggression could be defined as (1) the volume and duration of noise blasts delivered to another person, (2) the number of times a child punches an inflated toy clown, (3) the number of times a child fights with other children during recess, (4) a city’s homicide statistics gathered from police records, (5) a score on a personality measure of aggressiveness, (6) the number of times a pitcher hits a batter with a pitch during baseball games, and so on.

You may debate whether these different operationalizations (see [Chapter 5](#)) really measure hunger and aggression accurately. Different researchers will operationalize similar variables in different ways, depending on the research questions, the resources they have available, and their own

conception of the variable. In a research report, researchers must describe precisely how they operationalize all variables and also argue for why their operationalization is appropriate. These details of how a variable is operationalized also help other researchers to evaluate the study, and also potentially attempt to replicate the study.

Variables can be very broad and abstract, which can make operationalizing them a complex process involving many choices. A variable such as “word length” is rather narrow and concrete, and therefore easily operationalized in terms of numbers of letters or syllables in a word; although the exact words for the study must still be selected. In contrast, the concept of “stress” is very broad and more abstract. There are many different kinds of stressors—noise, crowding, relationship difficulties, familial expectations, academic workload, and so on. A researcher interested in stress will choose one stressor to study and then develop an operationalization for that specific stressor. The key point is that researchers must always translate variables into specific operations that will be manipulated or measured in a study.

Page 66



LO3 Avoiding Confounds in

Operationalizations

Ensuring that operationalizations are precise is crucial for making claims about these variables. If our operationalization captures more than just the variable of interest, it makes it difficult or impossible to interpret the results

of our research. Consider for example the variable “intelligence,” operationalized as “academic average.” You might argue that academic average captures more than just intelligence. It also reflects things like academic motivation, the personality trait conscientiousness, and perhaps even things like socio-economic status (e.g., wealthier people don’t have to work while enrolled, and therefore have more time to study). If intelligence, as well as the other things associated with academic average (like conscientiousness and motivation), are all related to our outcome variable (e.g., career success), then these other things are *confounding variables*, or simply confounds. Confounds are variables that we are not interested in, but are intertwined with our variable of interest, and can also explain our results. In the example above, if academic average predicts career success, we won’t know if this is due to intelligence or something else like conscientiousness. Confounds prevent us from knowing what variables are responsible for a given effect or association. As a result, they impede our ability to make claims. One way confounds occur in research (but not the only way) is when they are introduced by a lack of precision in our operationalizations (Greenland & Morgenstern, 2001; Weisberg, 2010). Confounds can occur for variables that are merely measured, but also in experiments when an independent variable is not manipulated in a precise way. When measuring variables, researchers must try to anticipate potential confounds and measure them as well, in order to rule them out statistically (although this can very difficult; cf., Westfall & Yarkoni, 2016). For example, when measuring the severity of schizophrenia symptoms, researchers might also measure past psychiatric assistance so they can rule out this potential confound (e.g., Sarlon, Millier, Aballéa, & Tourni, 2014).



Think about It!

Consider a study examining whether thinking about fast food makes people impatient (Zhong & DeVoe, 2010). In this study, thinking about fast food was the independent variable, manipulated by flashing pictures on a computer screen while participants did an unrelated task: either fast-food logos (e.g., McDonald’s, KFC) or same-sized squares. Can you spot the potential confound in this design? How would you change the design to

remove the confound? After thinking this through, you can read the original paper to find out how these researchers handled this confound in a second study. (Hint: If you need help finding this paper, check the references section of this book.)

Debating and Adapting Operationalizations

There is almost never a single, infallible method for operationalizing a variable. Many possibilities will be available, each of which has advantages and disadvantages. Researchers must decide which one to use based on the particular problem under study, the goals of the research, and other considerations such as ethics and cost. In some cases, advances in technology can change the operationalizations that are available to researchers (e.g., different forms of brain imaging). Researchers also sometimes disagree about whether an operationalization is an acceptable approximation of the variable of interest. Because so many possible operationalizations exist, and no one choice is perfect, our best understanding of a variable often involves studying it using a variety of different operationalizations. These definitions are fundamental to both non-experimental and experimental methods, the foundations of which we turn to next.

Non-experimental Method

Suppose a researcher is interested in the relationship between exercise and happiness. How could this topic be studied? Using a non-experimental method, the researcher could devise operationalizations for measuring both the amount of exercise that people engage in and their level of happiness. There is a variety of ways to operationalize both of these variables. For example, the researcher might simply ask people to report their exercise habits and current happiness (i.e., self-reports). The important point here is that for any non-experimental method, both variables of interest are measured and none are manipulated.

Now suppose that the researcher collects self-report data on exercise and happiness from many people and finds that exercise is positively related to happiness: The people who exercise more also have higher levels of happiness. The two variables would then be said to *covary* or *correlate* with each other. Variability (i.e., the different responses people give) in one variable (e.g., exercise) is associated with, or predicts, the variability in the other variable (e.g., happiness). In other words, differences in amount of exercise are associated with levels of happiness.

Because a typical non-experimental design allows us to detect covariation between variables, another term that is frequently used to denote some non-experiments is the *correlational method*. Correlational methods, methods in which variables are observed but not manipulated, should not be confused with the correlation statistic, which can be calculated for both experimental and non-experimental designs. In correlational methods, we examine whether the variables correlate or vary together. However, these designs do not allow for us to make inferences regarding causality or causal direction. However, they often provide a useful guide for future experimental research when possible, which can allow for causal inferences when properly conducted.



LO4 Relationships between

Variables

A great deal of research investigates the relationship between two variables. Are different values of one variable associated with different values of the other? That is, do the levels of the two variables vary together? As children grow older in age, do they also cooperate more with their playmates? If people are more depressed, are they also more likely to overeat?

Correlational designs allow us to consider many different kinds of relationships between variables. Recall that some variables have true numeric values (e.g., differences in people's height in centimetres), whereas the levels of other variables are simply different categories (e.g., nationality). For now, let's stick with variables that have a meaningful numeric value. (We discuss different kinds of variable scales further in [Chapter 5](#).) When we have two numeric variables, the relation between them can differ, taking the form of several different curves. Let's focus on the four most common relationships found in research: (1) the positive linear relationship, (2) the negative linear relationship, (3) the curvilinear relationship, and (4) when there is no relationship between the two variables. These relationships are best illustrated by line graphs that show how changes in one variable are accompanied by changes in the second variable. The four graphs in [Figure 4.1](#) show these four types of relationships (see also [Table 4.1](#)). We will also briefly consider another kind of relationship, the mediated relationship, which describes how the relationship between two variables can be explained via a third variable: This is often used to describe psychological processes.

Figure 4.1 Four types of relationships between variables

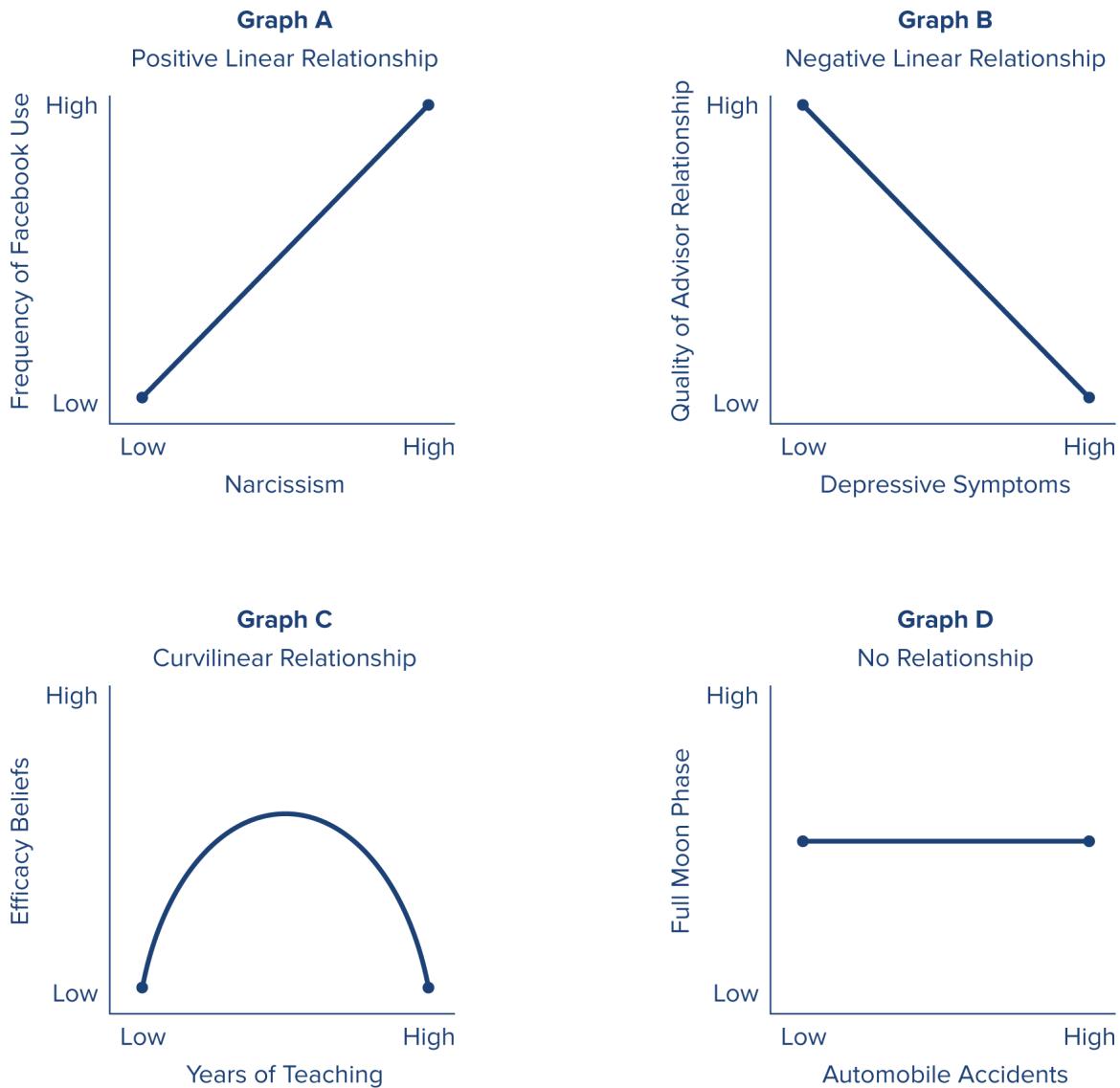


Table 4.1 Types of Relationship between Variables



TEST YOURSELF! Read the following examples and identify the type of relationship: positive, negative, or curvilinear.

Positive Negative Curvilinear

Positive Negative Curvilinear

1. Increased caloric intake is associated with increased body weight.
2. As people gain experience speaking in public, their anxiety level decreases.
3. The performance level of basketball players increases as arousal increases, from low to moderate levels, then decreases as arousal becomes too high.
4. Increased partying behaviour is associated with decreased grades.
5. Reducing the number of hours of sleep one gets is associated with a decrease in the ability to pay attention during class.
6. Amount of education is associated with higher income.
7. Liking a song increases the more you hear it, but then after a while you like it less the more you hear it.

Positive Negative Curvilinear

8. The more exercise your puppy gets, the less it chews on things.

Positive Linear Relationship

In a *positive linear relationship*, increases in the values of one variable are accompanied by increases in the values of the second variable. A common example is that age and weight have a positive linear relationship, especially for the first 20 years. From the moment you are born, the older you are, the heavier you are. For a more psychological example, consider a study conducted by Mehdizadeh (2010), in which students at York University reported their Facebook use and also completed a self-report measure of narcissism. Higher levels of narcissism are associated with holding a grandiose and overly positive view of the self; narcissists also tend to seek out admiration. These researchers found a positive linear relationship between how much someone used Facebook and their level of narcissism. The more narcissistic people were, the more time they spent on Facebook. Similarly, the more people spent on Facebook, the higher their narcissism score. To graph this relationship, we create a graph like the Test Yourself! box below, with a horizontal axis (going left to right, known as the *x*-axis) and a vertical axis (running top to bottom, known as the *y*-axis). Values of one variable are placed on the horizontal axis and run from low to high, from left to right. Values of the other variable are placed on the vertical axis and run from low to high, from top to bottom.

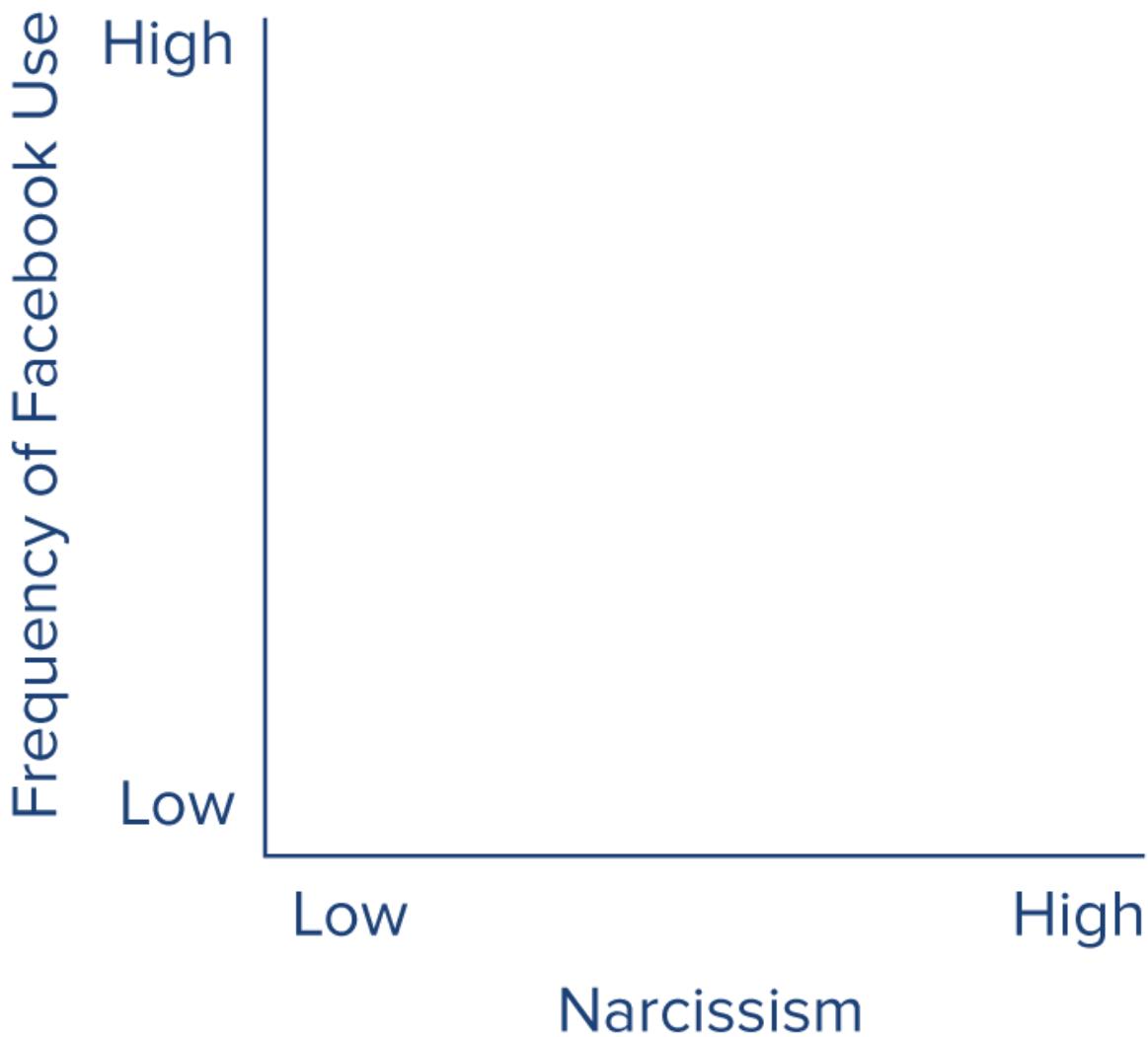


Test Yourself!

Try to imagine this relationship between Facebook use and narcissism, and draw a line on this graph that demonstrates how the two relate.

Graph A

Positive Linear Relationship



Negative Linear Relationship

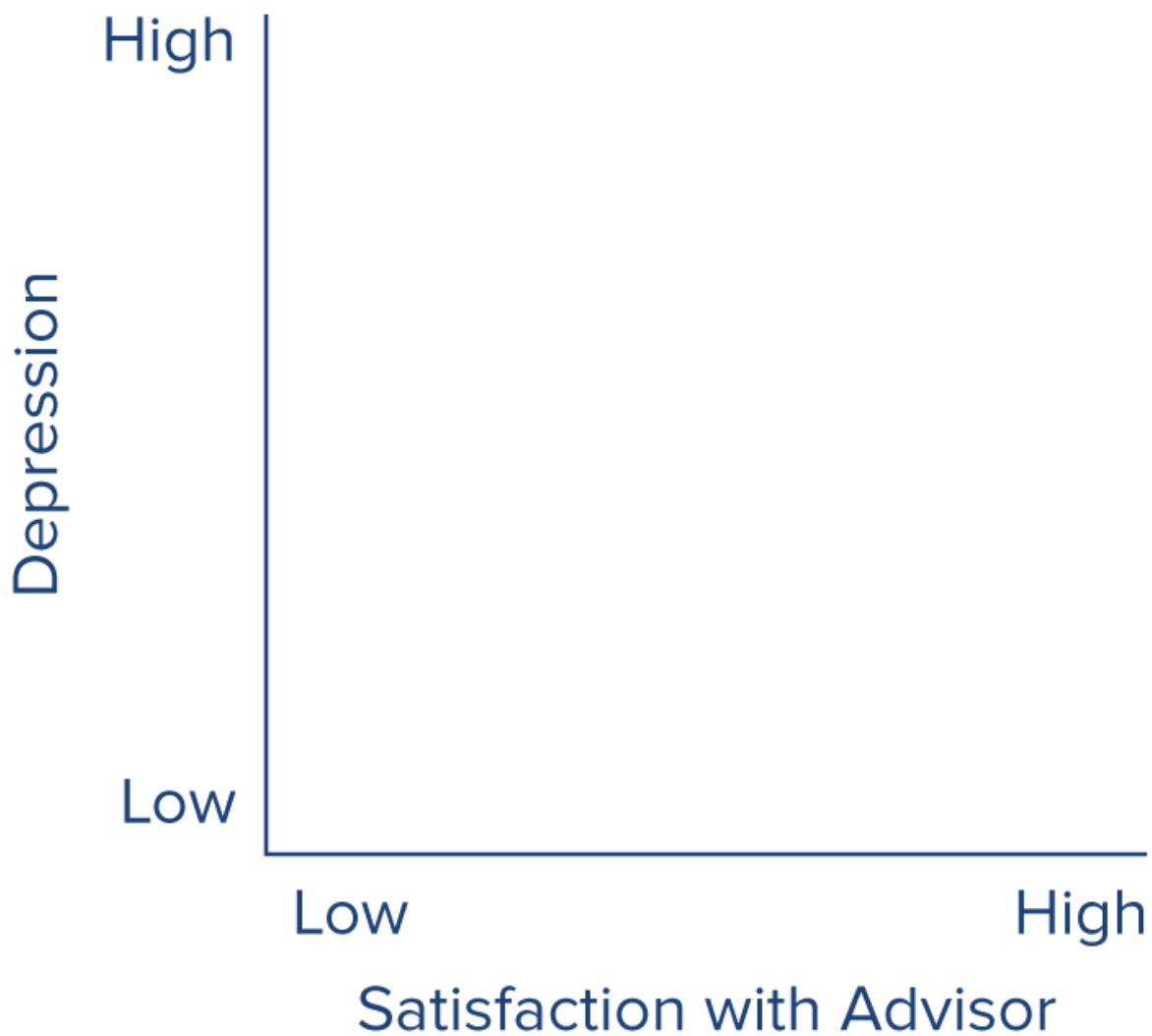
Variables can also be negatively related. In a [negative linear relationship](#), increases in the values of one variable are accompanied by decreases in the values of the other variable. So as values for one variable get bigger, the values for the other get smaller. As the amount of caffeine you consume increases, for example, the amount of sleep you can get decreases. Another example is a study by Peluso and colleagues (2011), who investigated depression among Canadian psychology graduate students. Graduate students completed a measure of

depressive symptoms, as well as questions about the relationship with their academic supervisor. As satisfaction with their advisor *increased*, the number of depressive symptoms that graduate students reported *decreased*. The two variables are systematically related, just as in a positive relationship, only the direction of the relationship here is reversed. Page 69



Test Yourself!

Try to imagine this relationship between satisfaction with advisor and depression, and draw a line on this graph that demonstrates how the two relate.



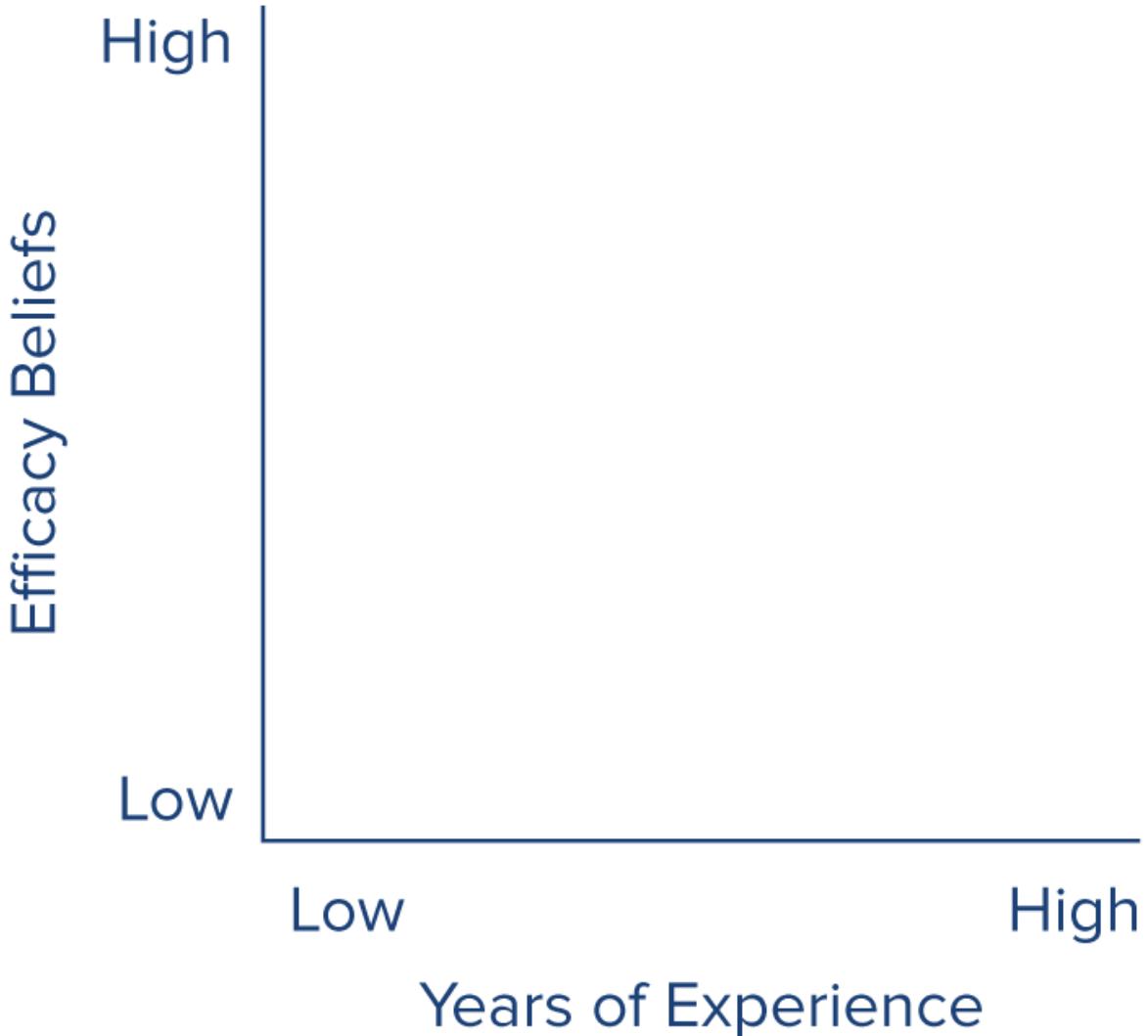
Curvilinear Relationship

In a *curvilinear relationship*, increases in the values of one variable are accompanied by both increases and decreases in the values of the other variable. In other words, the direction of the relationship changes at least once. For example, a relationship might begin being positive, but at a certain point it levels off and then becomes negative. Consider the relationship between eating Halloween candy and level of happiness. At the start, eating the candy might be associated with greater happiness (a positive relationship), but at a certain point if you keep eating and start to eat too much, eating more candy is going to be associated with less happiness (a negative relationship). Another example is the relationship between teachers' years of experience and their beliefs that they are effective at engaging students (Klassen & Chiu, 2010). Greater years of experience are initially accompanied by increases in believing they are effective at engaging students, but only up to a point. The relationship then becomes negative, as further increases in experience are accompanied by decreases in believing they are effective at engaging students. This particular relationship is called an *inverted-U relationship*.



Test Yourself!

Try to imagine this relationship between years of experience and efficacy beliefs, and draw a line on this graph that demonstrates how the two relate.



Page 70

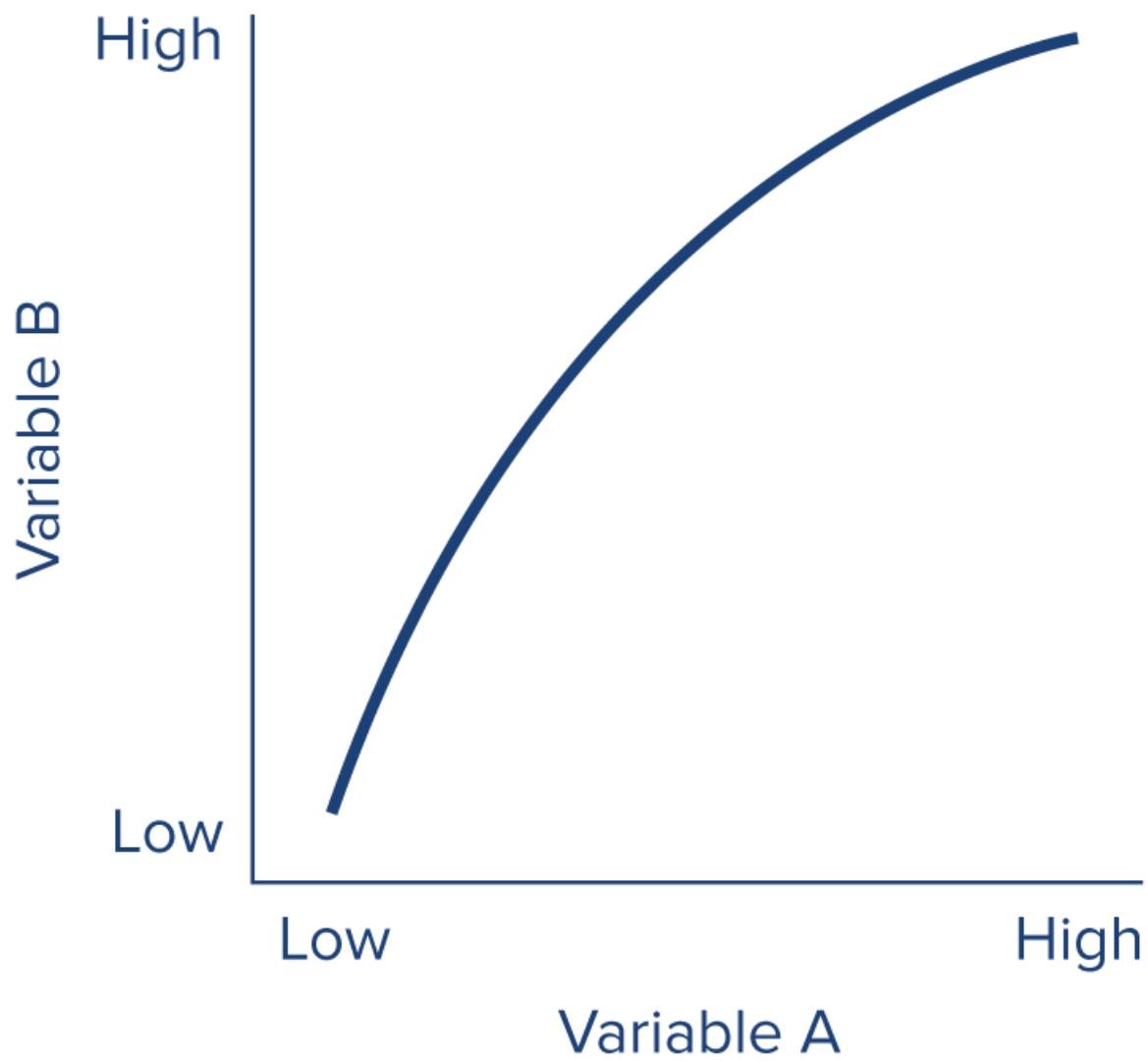
No Relationship

When there is absolutely no relationship between the two variables, the graph looks like a simple flat, horizontal line. Increases in the value of one variable have no relationship to the values of the other variable. This is what you find when you examine the relationship between automobile accidents and the phase of the moon (Laverty & Kelly, 1998). Unrelated variables vary independently of one another: Changes in one are not associated with changes in the other. Despite popular beliefs that a full moon can influence human behaviour, there is no evidence that lunar phase is associated with automobile accidents or other behaviours (Foster & Roenneberg, 2008).

Now check to see if you've completed all the above Test Yourself! exercises correctly by taking a look at the graphs in [Figure 4.1](#).

Page 71 Keep in mind that these graphs depict the most basic, clearest examples and are for illustration only. The relationship between two variables can result in almost any shape of line, with other relationships described by more complicated shapes than those in [Figure 4.1](#). For example, some variables are positively related to each other, but not in the strictly linear way we see in Graph A of [Figure 4.1](#). [Figure 4.2](#) depicts a positive relationship that is not strictly linear.

Figure 4.2 Positive non-linear function

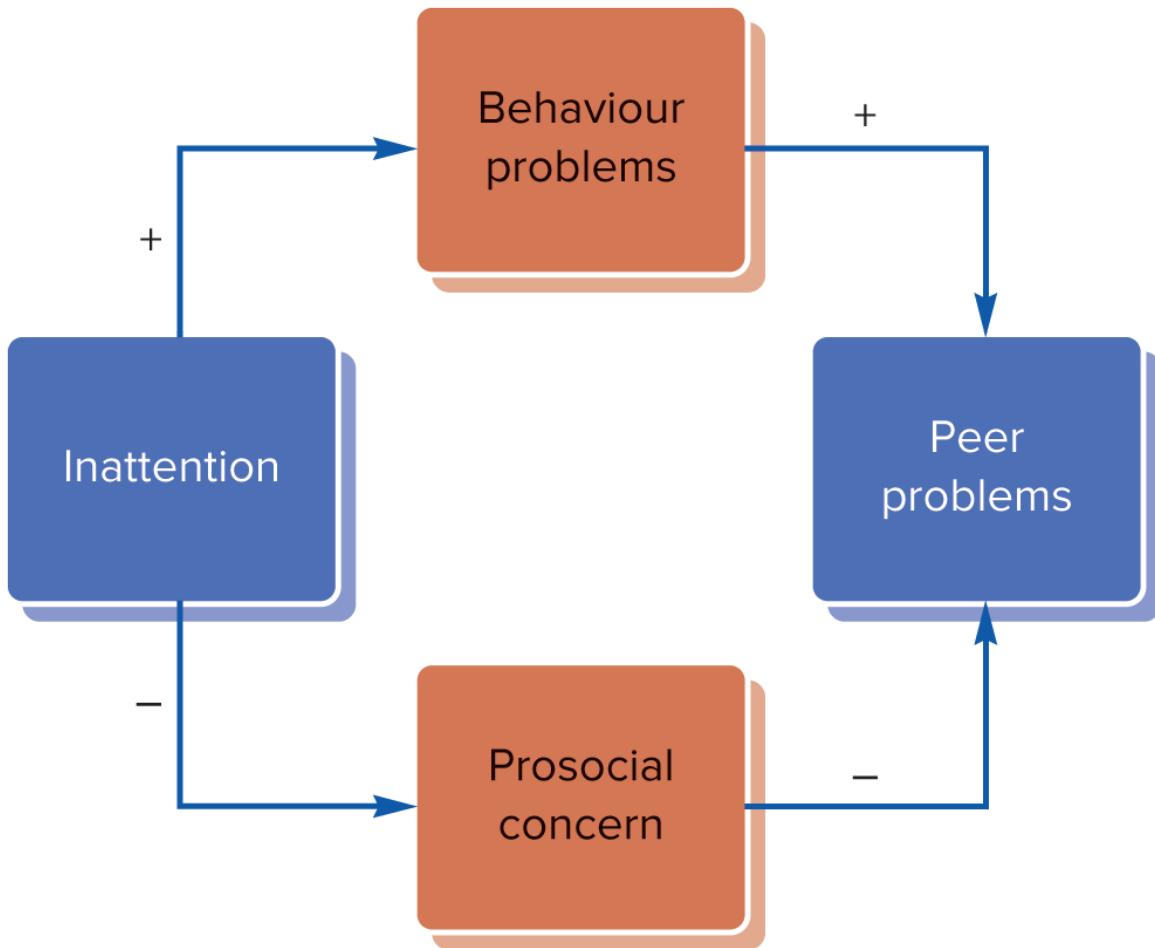


In addition to knowing the general type of relationship between two variables, it is also necessary to know the strength of the relationship. In non-experimental designs, this means we need to know the size of the correlation between the variables. Sometimes two variables are strongly related to each other, and there is little deviation from the straight lines depicted in [Figure 4.1](#), Graphs A and B. Other times the two variables are still correlated, but the association is weaker because scores deviate from the ideal, resulting in a cloud of points that don't all fall on a straight line. We can express the strength of a relationship between variables numerically using what is called a *correlation coefficient* (see [Chapter 12](#)).

Mediated Relationship

A [*mediating variable*](#) is a psychological process that helps to explain the relationship between two other variables (Baron & Kenny, 1986; Preacher & Hayes, 2004, 2008). Let's take the example of a study by Andrade and Tannock (2013), who found that the more trouble children had keeping their attention on something, the more problems they had relating to their peers. Two mediating variables were explored, which helped to explain this relationship between attention and relating to others (see [Figure 4.3](#)). The more inattentive the children were, the less they showed concern for their peers and the more behavioural problems they had, which then both related to more peer problems. In other words, a lack of concern for others and the presence of behaviour problems both mediated the relationship between inattention and peer problems. Models of mediated relationships like this one help provide insight into how and why variables are related to one another. With this knowledge, we can begin to explore possible interventions to help inattentive children make friends, by training them to direct more concern toward their peers, for example. Note that all four variables in this study could be described as participant variables, and all were measured using questionnaires completed by the children's teachers. These questionnaires are the researcher's operationalization for these constructs. A construct is simply another name for an abstract variable, idea, or phenomenon (e.g., social anxiety, happiness, impulsivity, generosity).Page 72

Figure 4.3 Prosocial concern and behaviour problems are mediating variables that help explain the relationship between inattention and peer problems.



Adapted from Andrade, B. F., & Tannock, R. (2013). The direct effects of inattention and hyperactivity/impulsivity on peer problems and mediating roles of prosocial and conduct problem behaviors in a community sample of children. *Journal of Attention Disorders*, 17, 670–680, Figure 1, p. 674.



Relationships between Variables Reduce Our Uncertainty

Relationships between variables are almost never perfect. A perfect relationship would mean that if we knew the value of one variable, we would know exactly what the value would be for the other variable. But in reality, things are much messier, and we're often only able to make a best guess about one variable, when we know the value of a related variable. For example, if we know that someone is very tall, we might be able to guess that that person is not very light, but we don't know his or her exact weight. Remember that the graphs in [Figure 4.1](#) depict

idealized patterns of relationships between two variables. Even if a positive linear relationship exists between two variables, it is highly unlikely that absolutely everyone who scores high on one will also score high on the other. Correlations capture the overall trend of relationships, averaging across many people's data. For this reason, individual deviations from the general pattern are likely. For example, although narcissism is generally related positively to frequency of Facebook activity (Buffardi & Campbell, 2008), not everyone who shows a high degree of Facebook activity will be highly narcissistic. There will be high and low narcissists who do not fit the general pattern. This is why you cannot use the results based on group data to make accurate guesses about any one particular individual, similar to how we shouldn't use stereotypes about groups to judge individuals. It is also why an individual case that doesn't fit the overall trend doesn't invalidate that group trend. Just because we know a person who smoked every day and never got cancer, this doesn't disprove the fact that, on average across a group, those who smoke are more likely to get cancer.

Another way to think about correlations is that we are reducing our uncertainty about the world by increasing our understanding of the variables we are examining. The term *uncertainty* means that there is something unknown or that we're not confident about knowing something. In some cases, we might ascribe an element of randomness to the unknown. Scientists refer to aspects of a phenomenon that we can't predict or explain, or that result from variables not of interest to us, as *random variability* or *error variability*. In general, a goal of research is to reduce random variability by identifying systematic relationships between variables. These relationships help to make things more predictable and therefore less uncertain. In terms of error variability, in the case of a survey measure for example, several influences not of interest will also contribute to the scores in addition to what we're actually interested in studying. These variables not of interest might include memory errors, a person misunderstanding a question, and inattentiveness, for example. Page 73

Correlations allow us to consider trends in association across many people's data. Even though error variability still exists, we will still be able to observe any overall trends. Relationships between variables are stronger when there is less error variability—when there is less noise in the data and more useful signal. For example, if 90 percent of high narcissists and only 10 percent of low narcissists used Facebook constantly, the relationship would be much stronger (with less uncertainty or randomness) than if 60 percent of high narcissists used Facebook constantly, but so do 40 percent of low narcissists.

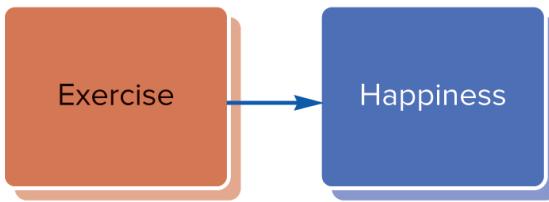
One way researchers seek to reduce random variability or error variability is to conduct additional research. With further studies, you may be able to identify other variables that are also related to Facebook use. For example, a Canadian study by Mehdizadeh (2010) found that having lower self-esteem in addition to higher narcissism also predicted greater Facebook use. Adding the variable of self-esteem to this analysis gave her an additional way to reduce uncertainty and increase prediction. But it is important to stress that we should always expect some uncertainty to remain: We're unlikely to be able to predict anything perfectly. We will revisit this idea of reducing error variability again in [Chapter 5](#) when we discuss reliability, and in [Chapter 12](#) when we discuss the correlation coefficient.

Interpreting the Results of Non-experimental designs

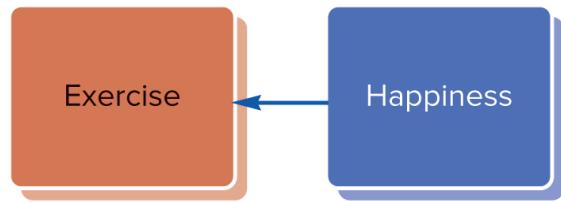
Non-experimental methods are a useful approach to studying relationships between variables. A relationship between variables means that the variables change together—the variables are said to covary or correlate with each other. Recall that this covariation is one of the three criteria for causal claims (see [Chapter 1](#)). However, non-experimental designs cannot tell us unequivocally whether two things are causally related. From these results, we can discover that two variables are related (i.e., they covary), but we cannot say for certain whether one has a causal impact on the other. Whenever two variables are related, say Variable A and Variable B, there are several possible explanations why we might observe this relationship. One is that A causes B. Another is that B actually causes A. And a third possibility is that some other variable that we haven't measured, C, may be why we observe this association between A and B; this is known as the *third-variable problem* (discussed below). In non-experimental designs, we often won't know which of these possibilities is behind the observed relationship (although this depends on the situation; in some cases it might be impossible for B to cause A). [Figure 4.4](#) illustrates these three possibilities. An important point to remember is that these three possibilities are not mutually exclusive. What this means is that if one is true, this doesn't necessarily preclude the others from also being true. In other words, it's entirely possible for all three to be true at the same time. Let's take a closer look at all these possible explanations for a correlation between two variables.

Figure 4.4 Possible causal directions in a hypothetical non-experimental study

Exercise causes increased happiness.

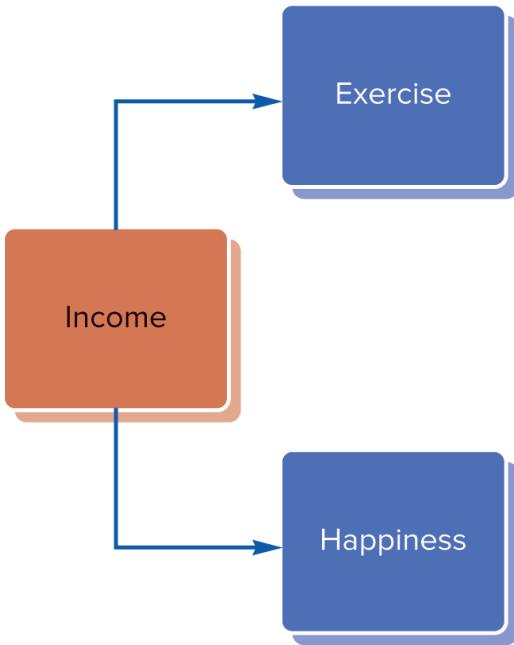


Happiness causes increased exercise.



A third variable such as income is associated with both variables, creating an apparent relationship between exercise and happiness.

Higher levels of income result in more exercise; higher income also leads to increased happiness.



Direction of Cause and Effect

With the non-experimental method, it is difficult to determine which variable causes the other, if both have causal effects on one another, or if there is no causal link between the two at all. Consider the example depicted in [Figure 4.4](#). After finding a positive relationship between exercise and happiness, we cannot conclude that exercise causes an increase in happiness. Although there are plausible reasons why exercise might cause people to be more happy, we must also consider the reasons why the pattern might be true. Perhaps people who are happier are more likely to be open to exercising. Or perhaps both possibilities are

true, and happiness and exercise have reciprocal causal influences on one another. One key issue to figure out what causes what is *temporal precedence*, or what comes first in time. This is another of the three criteria necessary to make causal inferences (see [Chapter 1](#)). If we could somehow gather information regarding temporal precedence, this would help us to figure out in what direction(s) the causal influence flows. Experimental designs can also be used to uncover the causal directionality of an association for some research questions. Importantly, knowing the direction of cause and effect has real-world implications. If exercise increases happiness, then starting an exercise program would be a reasonable way to increase happiness among a group of people. However, if increased happiness causes people to exercise more, simply forcing people to exercise may not increase their level of happiness.

The Third-Variable Problem

Another possibility is that the two variables we measured are not directly related to each other. In our example, exercise may not cause changes in happiness, and happiness may have no causal effect on exercise. Instead, a relationship between the two variables may be observed because some other variable that was not measured causes both exercise *and* happiness to increase. This is known as the [*third-variable problem*](#). Any number of other third variables may be responsible for an observed correlation between two variables. In the exercise and happiness example, one such third variable could be level of income. Perhaps being wealthier gives people the resources, and perhaps permits the free time, to get more exercise (e.g., the ability to afford a health club membership). Having more money could also increase levels of happiness. If income increases both exercise and happiness, it could (partially) explain why we see a positive association between these two variables, and there might or might not be a direct cause-and-effect relationship between exercise and happiness. Because third variables offer an alternative explanation for the observed relationship, correlations alone can never be considered evidence of cause. Recall from [Chapter 1](#) that the ability to eliminate alternative explanations for an observed relationship between two variables is another important factor when we try to discover whether one variable causes another.

Page 75

The third-variable problem is a different issue from the confounding variables discussed [earlier](#). Sometimes these terms have been used interchangeably (Greenland & Morgenstern, 2001), especially colloquially, but in psychology we distinguish between these two possible issues. Consider this example: Imagine you wanted to explore the relationship between weekly alcohol consumption and

academic grades. After asking people to report on both, you analyze the data and find a negative correlation: the more alcohol people consume per week, the lower their grades. What is the causal direction here? We do not know. It is possible that alcohol consumption causes low grades, and it is also possible that people with low grades try to drown their sorrows by drinking more alcohol. Or both might be true. Moreover, at least one third variable might be at work. For example, poor decision-making skills might cause people to drink more alcohol and also cause them to have low grades. Compare this example of a third variable to a potential confound. As a confound, what if people who drink more alcohol are also more socially active than people who drink less alcohol? Based on your operationalizations (i.e., the measurement of alcohol consumption but not social activity), you can't disentangle alcohol consumption from social activity. Thus, if you found a negative relationship between alcohol consumption and grades, you would not be able to tell whether this is strictly due to alcohol consumption, time spent socializing, or perhaps both. In other words, time spent socializing is confounded with alcohol consumption: the two cannot be separated. To summarize the difference here, third variables cause the apparent relationship between two other variables, whereas confounding variables are intertwined with another variable in your study so that you cannot tell which is at work.

As you can see, direction of cause and effect and potential third variables represent a limitation of non-experimental methods. However, these limitations are often not considered in media reports of research results. In fact, many media reports of non-experimental research mistakenly make causal conclusions.



TRY IT OUT!

Take a look at recent media articles reporting on a published empirical study. See if you can find an instance in which a non-experimental method has been used by the researchers, but the journalist makes conclusions regarding causality, or even just implies the presence of causality, based on these results. What sort of language is used to imply causality?

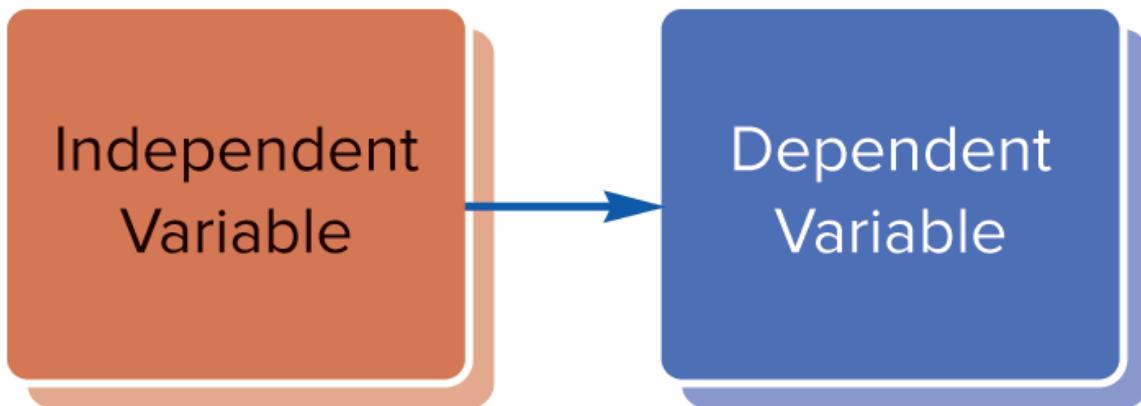
Only experimental methods can allow for causal inferences. If a research question is amenable to an experimental design, and the experiment is designed properly, then we can make causal inferences based on the results. We turn to experimental methods next to explain why this is the case.



LO5 Experimental Method

When researchers use the experimental method to investigate the relationship between variables, the *independent variable* is considered to be the “cause” and the *dependent variable* the “effect.” The independent variable is expected to cause changes in the dependent variable, and it is what we manipulate in an experiment. In contrast, the dependent variable represents the outcome, and it is only measured and not manipulated in an experiment. One way to remember these terms is to consider that the dependent variable *depends* on the level of the independent variable. Or another way to think about this is that the variable being manipulated in an experiment creates a situation that the participant has nothing to do with and has no control over: It’s something that the experimenter determines and is *independent* of the participant. In contrast, the dependent variable is the participant’s response to the manipulated variable (i.e., with the manipulated variable being the independent variable). Because the participant is responding to what happened in the experimental condition, the researcher assumes that what the individual does or says is caused by, or dependent on, the effect of the independent variable. For example, if we believe that exercise causes happiness, then we would design an experiment in which exercise is manipulated as the independent variable, and then we would measure levels of happiness as the dependent variable. You can visualize the causal relationship between the

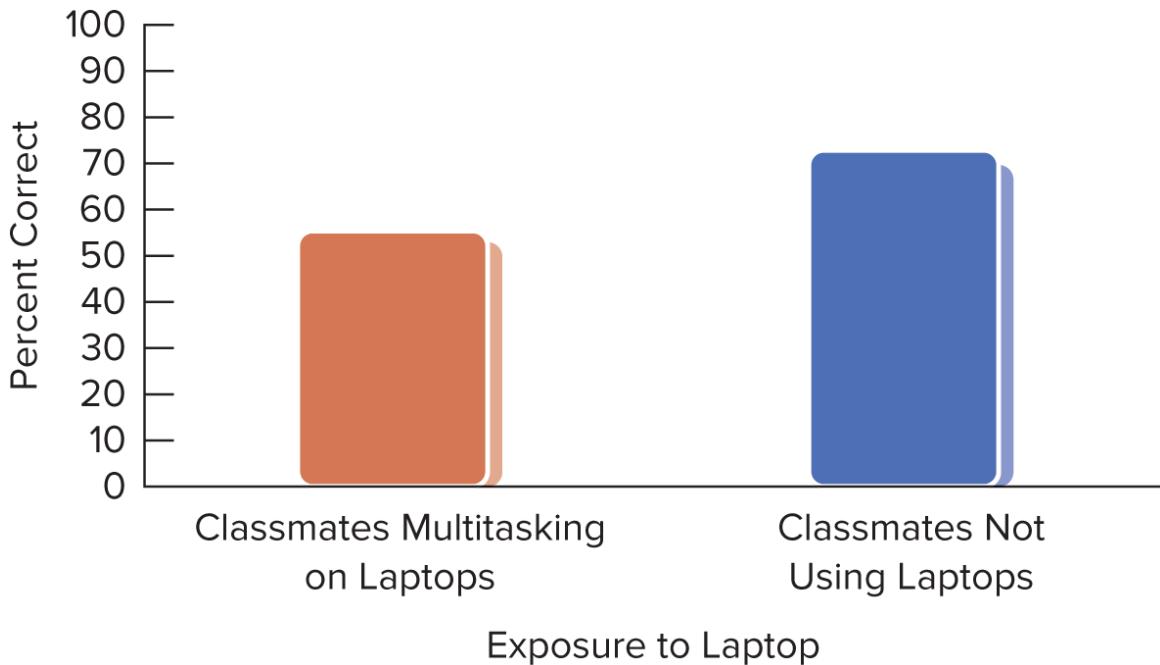
independent and dependent variables using an arrow, as we did in [Figure 4.4](#), with the arrow indicating the hypothesized direction of causation.



Let's consider an example to illustrate the features of an experiment here, as well as corresponding statistical analyses in [Chapter 13](#). Do you think that sitting behind someone multitasking on a laptop in class affects learning? A group of researchers from McMaster University and York University collaborated to conduct an experiment testing this research question (Sana, Weston, & Cepeda, 2013). The researchers hypothesized that exposure to the flickering laptop screen of a multitasking classmate (i.e., the independent variable) would decrease learning (i.e., the dependent variable). The independent variable was operationalized as having two levels or conditions, and participants were randomly assigned to experience one or the other. In the *experimental condition*, participants sat behind two confederates of the experimenter who were using their laptops to multitask during the lecture. In the *control condition*, participants sat behind two confederates who weren't using laptops. The dependent variable, learning, was operationalized as the percentage of correct answers on a comprehension test given right after the lecture. Participants who sat behind someone multitasking on a laptop scored 17 percent lower on the comprehension test than those who sat behind individuals without laptops. The researchers concluded that sitting behind flickering laptops impairs student learning.

When the relationship between an independent and a dependent variable is plotted in a graph, the independent variable is always placed on the horizontal axis (the *x*-axis) and the dependent variable is always placed on the vertical axis (the *y*-axis). [Figure 4.5](#) illustrates the results of this study in a graph. A bar graph is used because the experimental conditions are separate categories. (The graphs in [Figure 4.1](#) were line graphs because the variable on the horizontal axis is continuous; see [Chapter 12](#).)

Figure 4.5 Results of an experiment



Based on Sana, F., Weston, T., & Cepeda, N. J. (2013). Laptop multitasking hinders classroom learning for both users and nearby peers. *Computers & Education*, 62, 24–31, Figure 3.



Note that some research focuses primarily on the independent variable, with the researcher studying the effect of a single independent variable on numerous behaviours. Other researchers may focus on a specific dependent variable and study how various independent variables affect that one behaviour. Consider some possible experiments examining the relationship between group size and decision-making. One researcher studying this issue might focus on group size, varying the size of groups and looking at its effect on a wide variety of different behaviours (e.g., jury decisions, riskiness of group decisions). Another researcher might just be interested one kind of decision-making, such as jury decisions. In this case, the researcher could study how a wide range of different factors influence jury decisions (e.g., jury size, judge's instructions). Both approaches lead to important knowledge. [Figure 4.6](#) presents an opportunity to test your knowledge of the types of variables we have described.



Figure 4.6 *TEST YOURSELF!* Types of variables

Researchers conducted a study to examine the effect of music on exam scores. They hypothesized that scores would be higher when students listened to soft music compared to no music during the exam because the soft music would reduce students' test anxiety. One hundred (50 male, 50 female) students were randomly assigned to either the soft-music or no-music conditions. Students in the soft-music condition listened to music using headphones during the exam. Fifteen minutes after the exam began, researchers asked students to complete a questionnaire that measured test anxiety. Later, when the exams were completed and graded, the scores were recorded. As hypothesized, test anxiety was significantly lower and exam scores were significantly higher in the soft-music condition compared to the no-music condition.

Review the lists below and link each study variable to the type of variable it represents by drawing a line connecting the two.

Study variable	Type of variable
Sex of participant	Independent variable
Exam score	Participant variable
Headphones	Dependent variable
Music condition	Mediating variable
Test anxiety	Confounding variable



LO6 Designing Experiments That Allow for Causal Inferences

In the experimental method, the independent variable is manipulated and the dependent variable is then measured. Experiments allow for causal inferences in part because they incorporate temporal precedence. In many non-experimental designs, both variables are measured at the same time, so we don't have any information about what comes first. In an experiment, we manipulate the independent variable, and then we measure its effect on our dependent variable, so we have this temporal information. Page 78

Another aspect of experiments that helps to support causal inferences is the attempt to eliminate the influence of all other variables except for the one that is being manipulated (i.e., the one we're interested in as a cause). This control is usually achieved by trying to ensure that every other feature of the control and experimental conditions is held constant, except for the independent variable (which we are manipulating). If something cannot be held constant across conditions, we control for it by ensuring its effects are random, and therefore hopefully evenly distributed across conditions. For example, by randomly assigning people to our experimental or control conditions, the influence of extraneous variables associated with these participants that we're not interested in (e.g., participant variables) should be approximately equal across the conditions (provided there are enough participants). For example, whether someone comes into a study more or less motivated to pay attention cannot be controlled by the experimenter, but random assignment should balance out these effects roughly equally across the experimental and control conditions. Both procedures (i.e., control and random assignment) are used to try to eliminate alternative

explanations by ensuring that any differences observed between the conditions for the dependent variable are due to what's manipulated (i.e., the independent variable). Let's unpack these critical features of experiments in more detail.

Causality and Internal Validity

Experiments enable researchers to make causal claims about how variables are related to each other, when they are conducted properly. Recall from [Chapter 1](#) that inferences of cause and effect have three requirements. First, there must be *temporal precedence*: The causal variable must come first in the temporal order of events and then be followed by the effect. The experimental method addresses temporal order by first manipulating the independent variable and then observing whether it has an effect on the dependent variable. In other situations, you may observe the temporal order or you may logically conclude that one order is more plausible than another. In the study on laptop use during a lecture, laptop exposure came before the comprehension test (Sana et al., 2013).

The second requirement is that there must be *covariation* between the two variables: Changes in one variable must be accompanied by changes in the other. For an experiment, covariation is demonstrated when participants in an experimental condition exhibit a different effect relative to participants in a control condition. For the study on laptop use, those in a classroom where peers were multitasking on laptops (i.e., the experimental condition) had lower test scores relative to participants in classrooms where no one was using a laptop (i.e., the control condition) (Sana et al., 2013). In some experiments, covariation can be observed when the dependent variable is present or absent, depending on the level of the independent variable.

The third and final requirement for appropriate causal inference is the need to *eliminate plausible alternative explanations* for the observed relationship. Often, alternative explanations involve the possibility that some confounding variable could be responsible for the observed relationship. In the laptop-use study, you may have noticed that two different elements were present in the experimental condition but not in the control condition: a laptop and a multitasking screen. How can we know whether it is the multitasking screen of the laptop or just the laptop itself that affected participants' performance? Our inability to disentangle these two things is a potential confound that could provide an alternative explanation for these results. However, the researchers argue that this confound does not present an alternative explanation, but instead offers a realistic comparison that mimics actual classroom experiences (Sana et al., 2013). In the

real world, it would be very difficult to allow laptop use in a classroom, but somehow prevent people from doing other things with it other than take notes, so the point might be moot. If people have a laptop, they are likely going to multitask, such as checking their social media accounts in addition to taking notes. When designing research, a great deal of attention is paid to identifying and eliminating plausible alternative explanations, as this can be quite difficult. The experimental method attempts to rule out alternative explanations by using random assignment and different forms of experimental control (e.g., control conditions).Page 79

When a researcher designs an experiment that meets all three criteria for cause and effect, the experiment is said to have high internal validity. The term *validity* in research refers to concepts of both truth and accuracy, and there are many types of validity. *Internal validity* refers to the ability to draw accurate conclusions about causal relationships from an experiment's results. A study has high internal validity when it is well designed, so that the results support the inference that the independent variables caused changes in the dependent variable, with alternative explanations being implausible. Internal validity is achieved by designing high-quality experiments that are free from problematic issues, by using random assignment to condition and ensuring that only the independent variable of interest changes across conditions, for example.

Achieving Internal Validity: Experimental Control and Random Assignment to Condition

One way to eliminate alternative explanations is through *experimental control*, in which the only difference between conditions is the independent variable of interest, with everything else kept the same across conditions. In the laptop-use study, for participants in both the experimental and control conditions, the lecture that was presented was identical in both conditions and the only thing that differed was whether others were multitasking on a laptop (i.e., the experimental condition) or not (i.e., the control condition). Thus, when a difference was observed between conditions, it could be attributed to the presence or absence of a peer using a laptop to multitask.

If a variable is held constant across conditions, it cannot explain any changes in the dependent variable. Let's consider a potential experiment on whether exercise causes people to be happier. In this experiment, the research will have two conditions: (1) an experimental condition in which people participate in an hour-long boxercising class organized by the researchers (the exercise condition), and

(2) a control condition in which participants are not provided with such a class (the no-exercise condition). Importantly, the researcher wants to make sure that the only difference between the exercise and no-exercise groups is the exercise. If people in the exercise group are removed from their daily routine to engage in exercise, the people in the no-exercise group should also be removed from their daily routine to keep this aspect similar across conditions. Perhaps people in the no-exercise group could be asked to read a book for an hour instead of boxercising. If our control condition didn't include this element of a break, we wouldn't know whether any differences observed between conditions are due to the boxercising or simply having a break from one's normal routine.

To further explore this important issue, let's imagine what would happen if the no-exercise group is not given a break, like those in the exercise group. Suppose researchers find that those in the exercise condition have higher happiness levels than those in the no-exercise group. Now, instead of concluding that exercise caused increased happiness in this experiment, an alternative explanation exists: Perhaps simply having a break from routine causes increased happiness. After all, the no-exercise group did not benefit from such a break, and this could be why they maintained their existing levels of happiness. In this example, the break is a confounding variable: It is something that (1) differs between the experimental and control conditions, (2) is not a variable of interest to the researchers, and (3) could be an alternative explanation for the results observed. Another way to say this is that the break is a variable that is *confounded* with exercise. Let's consider another variation on this example. Imagine that everyone in the exercise condition spent the hour not only boxercising but also socializing with the others in the group, whereas everyone in the no-exercise condition spent their hour reading alone, not socializing with anyone else. Now we can see a different confound, the amount of socializing differs between conditions (along with exercise); it's not the variable of interest (i.e., it's not exercising), and it could provide a plausible alternative explanation in that socializing with others could easily make people happier. So it's now impossible to know whether it is exercise or socializing that results in differences in levels of happiness across conditions. Achieving appropriate levels of experimental control means avoiding such confounds and making sure only the independent variable of interest changes across conditions. As you can probably see, doing this is often very difficult and requires a lot of careful thought.

Page 80

Sometimes it is difficult, or even impossible, to keep a variable constant across all conditions. The most obvious case is any characteristic of the participants. Consider an experiment in which half the research participants choose to pay for

a boxercising class and the other half choose the no-exercise control condition. In this example, the participants in the two conditions might be different on some extraneous variable, such as income. Perhaps those who have more money are more likely to choose to pay for an exercise class compared to those who choose the control condition. If we later see greater happiness in the exercise condition, we don't know if it's because the people who chose this condition have a higher income or whether it's due to the actual amount of exercise being done. How can the researcher eliminate the influence of such extraneous variables in an experiment?

The experimental method attempts to eliminate the influence of such variables by random assignment to condition. Random assignment ensures that extraneous variables, variables not of interest to the researcher, are just as likely to affect one condition as they are to affect the others. This is more likely to hold true the more participants there are in the study, since randomness is more likely to result in equal distributions at higher samples (e.g., consider tossing a coin twice, versus 200 times). To eliminate the influence of individual characteristics (e.g., differences in participant motivation, attentiveness, rest, ability), the researcher assigns participants to the two conditions randomly. In actual practice, this means that assignment to groups is determined by a list randomizer, like the one that is available at www.random.org.



TRY IT OUT!

Imagine you are running an experiment and need to randomly assign 60 participants to either the experimental condition or the control condition. Using a word processor, create a list of the word “experimental” copied 30 times and the word “control” also copied 30 times. Then copy this entire list of 60 words into the list randomizer at www.random.org/lists, and click the Randomize button. The program will re-sort that list in random order. As each participant arrives to your study, you would use this randomized list to assign each participant to a condition.

By using random assignment along with a large sample, you can be relatively confident that many characteristics of the participants in the two groups will be virtually identical. For example, people with lower, medium, and higher incomes should be roughly equally distributed across the two groups.

Besides participant variables, many other variables that cannot be held constant are also controlled by random assignment. For instance, many experiments are conducted over a period of weeks, with participants being tested at various times on different days. Random assignment prevents a situation in which one condition is scheduled only during the mornings, with the other only taking place during the afternoons. If condition were confounded with time of day, then we would have difficulty inferring whether differences observed between conditions for the dependent variable are a result of the manipulated independent variable or simply the time of day.

Page 81

Experimental control and random assignment to condition help us to reduce the influence of confounding and extraneous variables. The goal of a good experimentalist is to achieve good control and randomization, so that an argument can be made that any difference between groups on the dependent (measured) variable can be attributed only to the influence of the independent (manipulated) variable. If this is the case, then the experiment has high internal validity.

Further Criteria for Claiming Cause

Let's review the three criteria for cause: (1) temporal precedence, (2) covariation, and (3) elimination of alternative explanations. Sometimes we impose even more stringent requirements before concluding that there is a causal relationship. Some philosophers, scientists, and students argue that cause-and-effect relationships are established only if the cause is both necessary *and* sufficient for the effect to occur. Suppose you have determined that reading the material for an exam is related to the grade you will get on the exam: Students who read the material score higher than students who do not read the material. To be *necessary*, the cause *must* be present for the effect to occur. In other words, you can't do well on the exam without reading the material being covered. To be *sufficient*, the cause will *always* produce the effect. This means that reading the material must always result in a high exam score.

Let's think about this a bit. If the exam is based only on material in the book, reading the book is probably *necessary* for a good grade on the exam. But is simply reading the material *sufficient* to do well? Or do other factors need to be considered, like making sure that you understand what it is that you are reading? Simply reading the material is not a sufficient cause in this example. Instead, you are most likely to retain the material when you actively study it by paying attention, relating the information to other things you know, and practise recalling

the material (for a review, see Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013).

It's important to note that the “necessary *and* sufficient” requirement for establishing cause is rare in psychology. Because most psychological outcomes have many different causes, it is extremely rare that just one cause is sufficient to explain some phenomenon. Many different things result in aggression, for example, or even laughter. Most behavioural scientists are not unduly concerned with the issue of ultimate cause and effect. Rather, they are more interested in carefully describing behaviour, studying how variables affect one another, and developing theories that explain behaviour. The general consensus is that there are few interesting “necessary and sufficient” causes of behaviour. Instead, research on numerous variables eventually leads to an understanding of a whole “causal network” in which a number of variables are involved in complex patterns of cause and effect.



LO7 Choosing a Method: Advantages of Multiple Methods

There are different advantages and disadvantages to both experimental and non-experimental designs. Researchers sometimes use both methods to test relationships among variables in different ways. Let's examine some of the issues that arise when choosing a method.

Artificiality of Experiments

In a laboratory experiment, the independent variable is manipulated within the carefully controlled confines of a laboratory. When done properly, this permits relatively unambiguous inferences concerning cause and effect, by reducing the possible influence of extraneous variables that aren't of interest. Laboratory experimentation is an extremely valuable way to study many problems. However, experiences in laboratory experiments can look quite different from real life, which raises questions about the degree to which we can learn about the real world from these results. In addition, not all research questions are amenable to an experimental design. There are many things we might be interested in that we wouldn't be able to easily

manipulate in a laboratory. Take your religious belief, for example. If you believe strongly in a religion, do you think someone could manipulate this to turn you into an atheist, using an experimental manipulation in the lab? Or what about your cultural background? Could a researcher change your cultural background in the lab? And yet religion and culture are two very important and interesting influences that we would like to study. Some research questions are more amenable to non-experimental research designs, and these designs can also sometimes better reflect the real world, reducing concerns about generalizability. It is important to keep in mind that all research aims to understand real-world phenomena, not things that only occur in a laboratory.

Page 82



Think about It!

In addition to religious belief and cultural background, what are some other interesting phenomenon that we think might have an important influence on our lives and experiences, but would likely be difficult or impossible to manipulate within the laboratory? Try to list as many of these as you can.

One way in which concerns about generalizability for experiments can be addressed is by attempting to conduct an experiment in a field setting. In a *field experiment*, the independent variable is manipulated in a natural setting out in the real world. As in any experiment, the researcher attempts to control extraneous variables via random assignment, experimental control, or both when possible. Many field experiments take place in public spaces such as street corners, shopping malls, or online. The advantage of a field experiment is that the investigation takes place in a natural context. The disadvantage is that the researcher loses the ability to directly control many aspects of the situation. A laboratory setting permits researchers to more easily control extraneous variables compared to a natural setting. But it is precisely this control that leads to the artificiality of the laboratory investigation, reducing its potential generalizability to the real world. Fortunately, when researchers have conducted experiments in both lab and field settings, the results of the experiments tend to be very similar (Anderson, Lindsay, & Bushman, 1999). However, in some cases,

researchers find just the opposite in the field compared to what they find in the lab, and the extent to which this occurs varies widely depending on the sub-field of psychology in question (Mitchell, 2012). In general, when only small effects were observed for laboratory studies, these were the findings for which opposite results were more likely to be observed out in the field (Mitchell, 2012).

☆ Student Spotlight: Field Experiments ☆

Spending time in nature has been found to have a positive impact on our psychological health, but most of this research has been conducted during warmer weather. Aeliesha Brooks, an undergraduate student at the University of Regina, was curious as to whether nature also had a positive impact during cold winter months (a great question for someone living in Saskatchewan!). Aeliesha and their collaborators conducted three separate experiments on the topic. Importantly, two of these experiments had a field component, in that people were randomly assigned to experience either a natural outdoor environment or a built indoor environment, with both experiences occurring out in the real world. Moreover, one of their experiments directly contrasted experiences out in the field with a more typical in-laboratory manipulation, the latter exposing in-lab participants to mere pictures of outdoor or indoor environments. As dependent variables, they included measures of positive and negative emotions, stress, anxiety, and depression, among others. You can learn about what they discovered in studies published in the *Journal of Environmental Psychology* (Brooks, Ottley, Arbuthnott, & Sevigny, 2017).

Page 83

Ethical and Practical Considerations

Sometimes the experimental method is not possible because experimentation would be either unethical or impractical. Consider child-rearing practices. Even if it were possible to randomly assign parents to one of two different child-rearing strategies, such as using physical types of punishment versus not, the manipulation would be highly unethical. Non-

experimental methods are crucial for understanding phenomena that we can't manipulate for ethical or practical reasons. Many important research areas present similar problems. Such problems need to be studied, and non-experimental designs can offer insights into relationships among these variables.



Think about It!

Revisit your list of interesting phenomena that would be difficult to manipulate in the laboratory. Can you add to this list, thinking now of additional things that it would be possible to manipulate, but unethical or impractical to do so?

When such variables are studied, people are often categorized into groups based on their experiences, forming our different conditions. When studying the effect of having an employed mother on child development, for example, one group would consist of children whose mothers work outside the home and another group would consist of children whose mothers do not work outside the home. In these types of studies, groups are formed on the basis of some actual real-world difference rather than through random assignment to a manipulation, as in an experiment. Because random assignment to condition is not possible, these designs cannot support causal inferences and are known as quasi-experimental designs. See [Chapter 10](#) for more information on quasi-experimental designs, which are used to compare groups when true experiments are impossible or impractical.

Describing Behaviour

As we discussed in [Chapter 1](#), describing a phenomenon is the primary goal of research: Without appropriate description, we cannot move forward with our other goals of predicting behaviour, uncovering causal influences, and explaining behaviour. When the research goal is to describe events accurately, causal claims are irrelevant and experiments are unnecessary. A great example of descriptive research in psychology comes from Jean

Piaget, who carefully observed the behaviour of his own children as they matured, and described in detail the changes in their ways of thinking about and responding to their environment (Piaget, 1952). Piaget's descriptions and his interpretations of his observations resulted in an important theory of cognitive development that triggered much experimental research and greatly increased our understanding of this topic (Flavell, 1996). Paul Rozin (2006, 2009) has written eloquently on the pitfalls a discipline encounters when it places a greater importance on explanation (e.g., through experimentation) relative to description. Page 84



Think about It!

How might seeking to explain a phenomenon before adequately describing it be problematic? Read the Rozin (2009) paper and compare your thoughts.

Predicting Future Behaviour

In many everyday situations, it can be helpful to successfully predict people's future behaviour—for example, success in school, ability to learn a new language, or likelihood of engaging in criminal activity. In such circumstances, issues of cause and effect may also be of little concern. It is possible to design measures that increase the accuracy of predicting future behaviour, without being concerned with ultimate causes. School counsellors can give tests to decide whether students should be in “enriched” classroom programs, employers can test applicants to help determine whether they should be hired, and forensic psychologists can evaluate convicted criminals to estimate their likelihood of committing another crime, all without being concerned about causes. These types of tests can lead to better decisions for many people. When researchers develop tests and questionnaires designed to predict future behaviour, they must conduct research to demonstrate that the measure does, in fact, relate to the behaviour in question (see [Chapter 5](#)).

Advantages of Multiple Methods

A complete understanding of any phenomenon requires research using multiple methods, both experimental and non-experimental. No method is perfect, and no single study is definitive. To illustrate, consider this research question: Do music lessons enhance intelligence? Some anecdotal evidence suggested yes, but empirical research findings yielded mixed results. Glenn Schellenberg, a researcher at the University of Toronto, set out to investigate. His first approach was to establish if there was a causal relationship between music lessons and improved intelligence (Schellenberg, 2004). Children entering Grade 1 were randomly assigned to one of four conditions: keyboard lessons, voice lessons, drama lessons, or no lessons. Thus, there were two music conditions (voice and keyboard) and two control conditions for comparison (drama and no lessons; note that these children received keyboard lessons the following year, known as a wait-list control). Change in intelligence was operationalized as the difference between intelligence scores before lessons versus after lessons, and the test used was a common, standardized measure designed for children. The hypothesis for this experiment was supported: Although intelligence scores for all groups increased over time, those children who received music lessons showed a greater boost in intelligence compared to either of the two control groups. From this, Dr. Schellenberg concluded that music lessons cause increases in intelligence. Page 85

Does the effect of music lessons on intelligence increase with more training? Does it last over time? These questions remained unanswered by the experiment, so Dr. Schellenberg (2006) switched to non-experimental methods. In one study, he asked first-year undergraduates to complete an intelligence test and to report on their history of music lessons and family background. As predicted, the length of time people had spent taking music lessons in childhood was positively related to their high school grades and their current intelligence scores in first year. You may be wondering if family income could be a confounding variable here. If people come from a high-income family, they may be more likely or able to take music lessons, and for longer, than people who come from a low-income family. Perhaps it is income, rather than music lessons, that predicts grades and intelligence scores. Dr. Schellenberg was also concerned about this possibility, so he

collected data on family income and removed the influence of this potential confound in the statistical analysis, finding the same result (see [Chapter 12](#) for details). Using a non-experimental method, Dr. Schellenberg thus showed that the length of music lessons in childhood is related to intelligence in early adulthood. We can't make a causal claims based on this study, but when coupled with the experiment reported in Schellenberg (2004), his research builds a case that music lessons directly impact intelligence. Note as well that establishing whether studying an instrument for many years has a greater influence on intelligence than only studying for a few years would have been very difficult to do with an experiment. It would be pretty expensive, and perhaps even impossible, to randomly assign children to a decade or more of violin lessons!

The important message here is that no research method is a perfect test of a hypothesis. There are a great many methods available, and every method has some strengths and some weaknesses. This is why doing multiple studies using multiple different methods is often the best way to proceed: The strengths of one method might well compensate for the weaknesses in another. And once different studies using different methods all point to the same conclusion, our confidence in the findings and our understanding of the phenomenon are greatly increased.

In the remainder of this book, many different methods will be discussed, all of which are useful under different circumstances. In fact, all are necessary to understand the wide variety of behaviours that are of interest to behavioural scientists. Complete understanding of any issue requires research using a variety of methodological approaches.



Illustrative Article: Studying Behaviour

Many people have had the experience of anticipating something bad happening to them: “I’m not going to get that job” or “I’m going to fail this test” or “She’ll laugh in my face if I ask her out!” Do you think that anticipating a negative outcome means that a person is less distressed when

a negative outcome occurs? That is, is it better to think “I’m going to fail” if, indeed, you may fail?

In a study published by Golub, Gilbert, and Wilson (2009), two experiments and a field study were conducted in an effort to determine whether this negative expectation is a good thing or a bad thing.

In the two laboratory studies, participants were asked to complete a personality assessment and were then led to have either positive, negative, or no expectations about the results. Participants’ affective (emotional) state was assessed prior to—and directly after—hearing a negative (in the case of study 1a) or positive (in the case of study 1b) outcome. In the field study, participants were undergraduate introductory psychology students who were asked about their expectations of their performance in an upcoming exam. Then, a day after the exam, positive and negative emotion were assessed. Taken together, the results of these three studies suggest that anticipating bad outcomes may be an ineffective path to positive emotion.

First, acquire and read the article:

- Golub, S. A., Gilbert, D. T., & Wilson, T. D. (2009). Anticipating one’s troubles: The costs and benefits of negative expectations. *Emotion*, 9, 227–281. doi:10.1037/a0014716

Page 86 Then, after reading the article, consider the following:

1. For each of the studies, how did Golub, Gilbert, and Wilson (2009) operationalize the *positive expectations*? How did they operationalize *affect*?
2. In experiments 1a and 1b, what were the independent variable(s)? What where the dependent variable(s)?
3. This article includes three different studies. In this case, what are the advantages to answering the research question using multiple methods?

4. On what basis did the authors conclude, “Our studies suggest that the affective benefits of negative expectations may be more elusive than their costs” (p. 280)?
5. Evaluate the external validity of the two experiments and one field study that Golub, Gilbert, and Wilson (2009) conducted.
6. How good was the internal validity?

Study Terms

Test yourself! Define and generate an example of each of these key terms.

- confounding variables (p. 66).
- curvilinear relationship (p. 69).
- dependent variable (p. 76).
- experimental control (p. 79).
- experimental method (p. 64).
- field experiment (p. 82).
- independent variable (p. 76).
- internal validity (p. 79).
- mediating variable (p. 71).
- negative linear relationship (p. 68).
- non-experimental method (p. 63).
- operationalization (p. 65).
- participant variable (p. 65).
- positive linear relationship (p. 68).
- random assignment (p. 80).
- response variable (p. 65).
- situational variable (p. 65).

- *third-variable problem* (p. 74).
- *variable* (p. 63).

Review Questions

Test yourself on this chapter's learning objectives. Can you answer each of these questions?

1. What are the major differences between non-experimental and experimental research methods?
2. What is an operationalization of a variable? List three variables and give at least two possible operationalizations for each.
3. How do confounds relate to operationalizations? Are any of your operationalizations for question 2 confounded with another variable? Can you change your operationalization to eliminate the confound?
Page 87
4. Distinguish between positive linear, negative linear, and curvilinear relationships.
5. How do causal direction and third variables qualify our interpretation of correlations?
6. What is the difference between an independent variable and a dependent variable?
7. How do experimental control and random assignment influence causal claims?
8. Describe the three elements required for inferring causation. Which one(s) are achieved by experimental methods? By non-experimental methods? How are they achieved?
9. What are some advantages and disadvantages of the two basic research approaches?

Deepen Your Understanding

Develop your mastery of these concepts by considering these application questions. Compare your responses with those from other people in your study group.

1. Researchers have found that elementary school students who sit at the front of the classroom tend to get higher grades than students who sit at the back of the classroom (Tagliacollo, Volpato, & Pereira, 2010). What are three possible cause-and-effect relationships for this non-experimental result?
2. Consider the hypothesis that work stress causes family conflict.
 1. How might you investigate the hypothesis using the experimental method? How would you operationalize each variable?
 2. Identify the independent variable and the dependent variable.
 3. How could you translate this experimental hypothesis into a non-experimental hypothesis? What type of relationship between stress at work and family conflict at home is proposed (e.g., positive linear, negative linear)?
 4. Graph the proposed relationship.
 5. How might you investigate your non-experimental hypothesis? How would you operationalize each variable?
 6. Would you prefer to use the experimental or non-experimental method to study the relationship between work stress and family conflict? What are the strengths and weaknesses of each approach?
3. Identify the independent and dependent variables in the following descriptions of experiments. Once you have identified the independent

variable, also identify its different levels. Can you spot any confounds?

1. Students watched a cartoon either alone or with others and then rated how funny they found the cartoon.
2. A comprehension test was given to students after they had studied textbook material either in silence or with the television turned on.
3. Some elementary school teachers were told that a child's parents were university graduates, and other teachers were told that the child's parents had not finished high school; they then rated the child's academic potential.
4. Workers at a company were assigned to one of two conditions: One group of workers completed a stress-management training program, another group of workers did not participate in any training. For the next two months, the number of sick days that workers took was recorded.
Page 88
4. The limitations of non-experimental research were highlighted by the results of an experiment on the effects of postmenopausal hormone replacement therapy. In the experiment, participants were randomly assigned to receive either hormone replacement therapy or a placebo control (i.e., no hormones). In 2002, the investigators concluded that women taking hormone replacement therapy had a higher incidence of heart disease than did women in the control condition. At that point, they stopped the experiment and informed both the participants and the public that they should talk with their physicians about the advisability of this therapy. This finding was the opposite of what had been found in non-experimental research, in which women taking hormones had a lower incidence of heart disease. In these non-experimental studies, researchers compared women who were already taking the hormones with women not taking hormones. Why do you think the results were different between the experimental research and the non-experimental research?

Chapter 5

Page 89

Measurement



©Floridapfe from S.Korea Kim in cherl/Getty Images

Researchers sometimes wish to measure finger length as an indirect measure of pre-natal hormone exposure, but is this tiny monkey using the best measurement approach?

Learning Objectives

Keep these learning objectives in mind as you read to help you identify the most critical information in this chapter.

By the end of this chapter, you should be able to:

1. [LO1](#) Define the reliability of a measure and describe the differences among test-retest, internal consistency, and inter-rater reliability.
2. [LO2](#) Compare and contrast reliability and validity.
3. [LO3](#) Describe how a researcher can build a case for construct validity, including predictive validity, concurrent validity, convergent validity, and discriminant validity.
4. [LO4](#) Describe the problem of reactivity for a measure and discuss ways to minimize reactivity.
5. [LO5](#) Describe the properties of the four scales of measurement: nominal, ordinal, interval, and ratio.

Page 90 Every variable that is studied must be operationalized. Recall from [Chapter 4](#) that operationalizations are the specific methods used to manipulate or measure a variable. Imagine that you are interested in studying the degree to which relationship satisfaction predicts whether a relationship will end.

Operationalizing the end of a relationship requires that you specify a time frame that you will study: Will you examine relationship status after a few days, months, or years? And measuring relationship satisfaction could involve simply asking people how satisfied they are with their romantic relationship, using a response scale that ranges from 1 (*not at all satisfied*) to 9 (*very satisfied*). In fact, LeBel and Campbell (2009) at Western University used this sort of self-report measure for relationship satisfaction and found that it was able to predict whether people were still in that romantic relationship after four months. Because you can operationalize variables in many different ways, we will focus on the primary

criteria researchers use to evaluate operationalizations: reliability and validity. We will also consider some other characteristics that are important when operationalizing variables, including reactivity and scales of measurement.

Self-Report Measures

In this chapter, much of the discussion of reliability and validity will refer to self-report measures. Although reliability and validity are important characteristics of all operationalizations, systematic research on reliability and validity is often carried out on self-report measures of individual differences. Psychologists often use self-report measures to study a wide variety of topics, including attitudes, evaluations, motivations, and preferences. Self-report measures are also often used to study personality traits, which are relatively stable tendencies in how people think and feel. For example, Costa and McCrae (1985) developed the *NEO Personality Inventory* (NEO-PI) to measure five broad personality traits: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism (often labelled as its inverse, emotional stability). Other self-report measures are important for applied settings. Clinical, counselling, and personnel psychologists use scales to help make better clinical diagnoses (e.g., *MMPI-II*; Hathaway, McKinley, & MMPI Restandardization Committee, 1989), career-choice decisions (e.g., the *Strong Interest Inventory*; Harmon, DeWitt, Campbell, & Hansen, 1994), and hiring decisions.

When doing research in these areas, it is usually wise to use existing measures rather than develop your own, if possible. Existing measures should have reliability and validity data to help you decide which measure to use, and creating your own measure takes a lot of work (primarily to establish its reliability and validity). Using an existing measure means you will also be able to compare your findings with prior research using the same measure. Some existing measures are owned and distributed by test publishers and are primarily used by professional psychologists in applied settings (e.g., schools and clinical practices). Other measures are freely available for researchers to use in their own research. But how do you find these measures? One source is the *Mental Measurements Yearbook*, which is available as a database from your library (similar to *PsycINFO*). You can also find measures by searching *PsycINFO*, and even with a basic Internet search. Choosing from among many different possible measures can be

confusing. But understanding the concepts of reliability and validity will help you to evaluate the quality of existing measures.



LO1 Reliability

One of the most important aspects of a good operationalization is *reliability*, which refers to the consistency or stability of a measure. A highly reliable measure will give you the about the same result, every time you use it to measure the same thing. Imagine using a fancy new feature on your smartphone that measures the length of things using its camera. If you're measuring the same object, then it should give you the same estimate of length, and we would then consider this to be a reliable measurement. But if you're measuring the same object and getting a different estimate of length every time, we'd consider this to be an unreliable measure, and perhaps you'd be better off using an old wooden ruler.

Similarly, a highly reliable measure of a relatively stable psychological variable such as conscientiousness will yield about the same result each time you administer the test to the same person. The test would be unreliable if it found someone to have an average level of conscientiousness at first, then a low level on the next administration, and then a high level upon a third administration. Put simply, a reliable measure does not fluctuate much from one measurement instance to the next when measuring the same thing. Fluctuations in measurement can be attributed to error in the measurement tool. Note, however, that *all measures have some error*: No measure is perfect and free from variability in scores not associated with the central construct of interest.

A more formal way to understand reliability involves two concepts: true score and measurement error. Any measurement includes two components: (1) a *true*

[score](#), which is the person's actual level of the variable of interest (*not* the score they get on the measure of that variable), and (2) [measurement error](#), which is any contributor to a measure's score that is not based on the actual level of the variable of interest (i.e., not the true score). So, for example, imagine we're measuring how quickly people can react to a stimulus (i.e., their reaction time) by having them press a button on a keyboard as soon as they hear a tone. Our reaction time measure is calculated as the number of milliseconds it takes from the onset of the tone until people press a button. This time, between tone and button-press, is determined most of all by people's reaction speed. However, other factors might also play a role. Imagine that someone reacts quickly but presses the wrong button, and then that person has to correct and press the right button. This extra delay that comes about from having pressed the wrong button is measurement error: It's something that contributes to our measure of reaction time, but it's not related to what we're truly interested in (how fast they can react) and instead is something else (e.g., accuracy of finger presses). All measures contain measurement error, and this measurement error leads to less reliable measure. The key factor to consider is *how much* measurement error exists in a measure. Too much measurement error, and not even measurement of the true score, makes for a very unreliable measure that gives you a very different result every time. Unreliable measures also don't give you a good indication of the true state of affairs. Recall the smartphone function that gives a different estimate of length for the same object every time you use it: How could this possibly tell you something accurately about this object? In contrast, measures that are very reliable have less measurement error, proportionally speaking: The scores on the measure derive much more from the true score than from measurement error. These measures will also yield very similar (or nearly identical) scores on repeated administrations. When developing a scale, improving the reliability of a measure reduces uncertainty or measurement error captured by a measure's score (see [Chapter 4](#)).

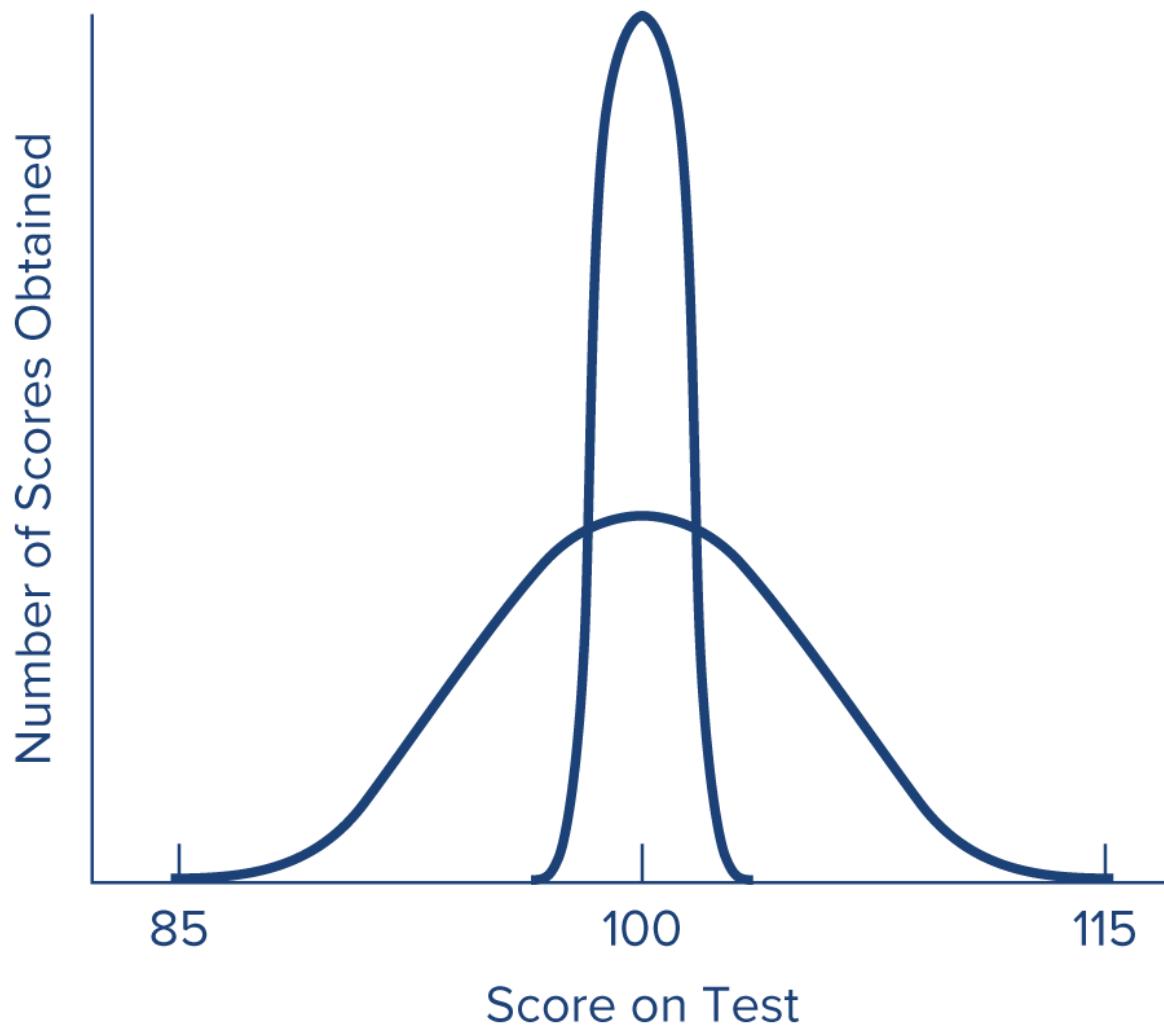


Think about It!

Imagine that you know that someone's true level of some stable psychological variable (i.e., their true score) is equivalent to a score of 100 on two different tests. Now suppose that you administer these two different tests to that person each week for a year. You are curious to know which test is the more reliable one. After the year, you calculate the person's average score on each test, across the 52 scores you obtained. What might your data look like? Hypothetical data are

shown in [Figure 5.1](#). For each test, the average score is 100, which is centred on the person's true level of this variable. However, scores on one test range from 85 to 115 (a 30-point spread), whereas scores on the other test range from 97 to 103 (a 6-point spread). Which test is more reliable?

Figure 5.1 Comparing data for a reliable measure with an unreliable measure



When conducting research, you often can only measure each person once. Thus, it is very important that you use a reliable measure. To the extent that your measure is reliable (and valid), your single administration of the measure will reflect the person's true score.

Reliability is crucial for any measurement. If our measure is not reliable, then we don't really know what the scores on this measure indicate. Moreover, all that excess measurement error will make it hard to detect any true relationship between variables, because the amount of true variance being captured (i.e., the true score) will be small and masked by all the noisy and random measurement error. Trying to study behaviour using unreliable measures is a waste of time because conclusions will be meaningless and the results will not be replicable. Keep in mind, however, that all measures contain some amount of measurement error. The key is to try to minimize this measurement error and maximize the amount of true score being captured by any particular measure. In truth, it's probably more accurate to talk about measures having high reliability or low reliability, rather than being totally reliable or totally unreliable, per se.

Researchers strive for reliability in their measures in different ways, depending on the type of operationalization and measure being employed. When observing behaviour in the real world, increasing reliability for a measurement might entail carefully training observers to notice and record behaviour in a highly systematic way. This careful training helps to maximize the true score and reduce the measurement error that might emerge from different observers coding things differently. For self-report measures, increasing reliability would typically mean paying close attention to the way questions are phrased, to avoid confusion in participants. In a psychophysiological study, this could mean being careful and systematic in how recording electrodes are placed on the body to measure physiological reactions. In many areas, reliability can be increased by making multiple observations of the same variable. This is akin to the old carpenter's adage: measure twice, cut once. This approach of measuring something more than once is most commonly seen when assessing personality traits and cognitive abilities. A personality scale, for example, will typically have ten or more questions (i.e., items) designed to assess the same trait, and you can think of each item as a new attempt at measuring the same construct. Reliability typically increases as the number of items increases.



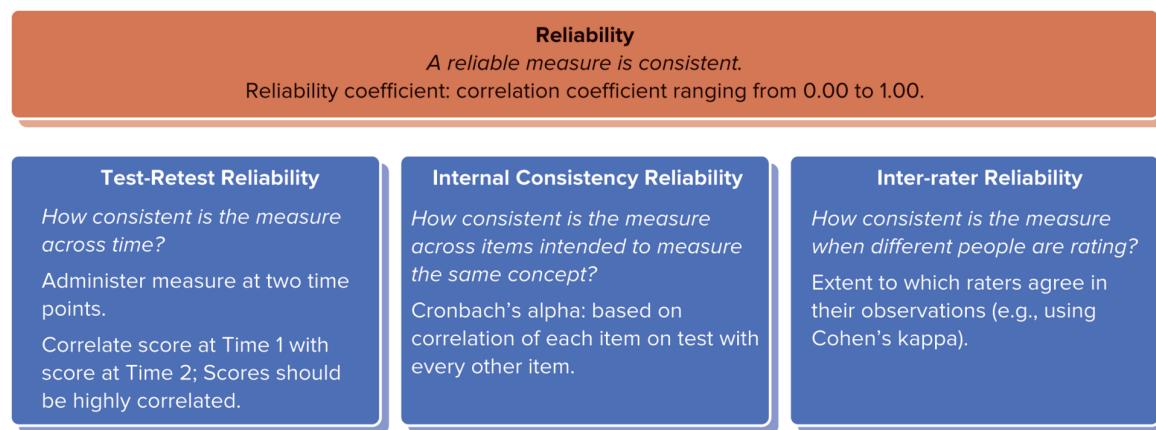
Think about It!

All the examples above attempt to increase the reliability of measurement by decreasing measurement error, but in different ways. Go through each example and try to identify the specific source of measurement error and how it is being

minimized. Think of other forms of measurement and try to imagine what some sources of measurement error might be and how it might be minimized.

How can we know how reliable a measure is? Unfortunately, we cannot directly observe the true score and error components of a score on some measure. However, we can assess the stability of measures using a correlation coefficient, and this allows us to investigate different types of reliability for a measure. Recall from [Chapter 4](#) that a correlation coefficient is a number that tells us how strongly two variables are related to each other—the degree to which changes in one score are accompanied by changes in another score. There are several ways of calculating correlation coefficients, with the most common being the [*Pearson correlation coefficient*](#). Pearson correlations range from 0.00 to +1.00 when positive, and 0.00 to -1.00 when negative. The mathematical symbol for the Pearson correlation is a lowercase r . A correlation of 0.00 tells us that the two variables are completely unrelated, and the closer a correlation is to either +1.00 or -1.00, the stronger the relationship. When the correlation is positive (a “plus” sign), there is a positive linear relationship: Increases in one variable accompany increases in another. A negative linear relationship is indicated by a “minus” sign: Increases in one variable accompany decreases in another. The Pearson correlation coefficient will be discussed further in [Chapter 12](#). When you read about reliability, correlations are sometimes referred to as *reliability coefficients*. There are several different types of reliability for a measure that can be investigated using the correlation coefficient, with [Figure 5.2](#) providing an overview.

Figure 5.2 Indicators of reliability of a measure



Test-Retest Reliability

Test-retest reliability is assessed by giving many people the same measure twice. For example, the reliability of a test could be assessed by giving it to a group of people on one day and then again a week later. With two scores for each person, we would calculate a correlation coefficient to determine the relationship between the first test score and the second test score (i.e., the “retest”). If there is a strong positive correlation between the two sets of scores, this would mean that those who score high on the first administration also score high on the second, indicating a high level of reliability. As with most things in science, there is no strict and agreed-upon cut-off for determining when this correlation is high enough to be acceptable, but some guidelines suggest a test-retest correlation of at least .80. This might differ depending on the particular measure, context, and research being conducted.

There are variations of test-retest reliability that can be useful in some circumstances. For example, because the test-retest approach involves administering the same test twice, correlations between the sets of scores could be inflated if people are able to remember how they responded the first time. This might be especially likely with a short test, with memorable questions, and with a short gap between the two administrations. One option to address this concern is *alternate forms reliability*, in which two different forms (or versions) of the same test are administered on two separate occasions.

Some psychological constructs are relatively stable across time, such as personality traits. For these constructs, we would expect test-retest reliability to be relatively high. However, other variables are less stable and expected to change from one moment to the next. Our current mood, for example, might easily be different at different times of our life. You are probably a lot grumpier during the exam period compared to the summer months. In this case, the test-retest reliability for a measure of current mood would not be expected to be high, and might not even be the most appropriate form of reliability to assess. Another factor to take into account is that obtaining two measures from the same people at different points in time can be difficult. This is why this form of reliability is often only investigated by researchers in the process of developing a scale, who are invested in demonstrating the reliability and validity of their new measure.

Internal Consistency Reliability

Internal consistency reliability examines how successful the different items in a scale are at measuring the same construct or variable. Think of each item as a different attempt to measure that same construct. When people respond similarly across these different attempts (i.e., different items), it suggests that the measure is reliable. Although the number of items in a measure varies across different scales, a person's test score is typically based on the total of their responses on all items, or an average across all items. This is just like how your grade on a test is the total of all your correct answers for the many questions on the test.

Perhaps the most common indicator of internal consistency is a statistic called Cronbach's alpha (α). In this analysis, the researcher calculates how well each item correlates with every other item, which produces a large number of inter-item correlations. The value of Cronbach's alpha is based on the average of all inter-item correlations and the number of items in the measure. It is also possible to examine the correlation between each item and the total score based on all items. These item-total correlations provide information about each individual item and its relation to the total score. Items that do not correlate with the others can be eliminated to increase the measure's reliability. This is also useful when trying to construct a briefer version of a measure. Even though reliability can increase with longer measures, a shorter version can be more convenient to administer and also have acceptable reliability. Although Cronbach's alpha is the most often used indicator of internal consistency reliability, it is not without its problems (Schmitt, 1996). Recently, a superior alternative to Cronbach's alpha has emerged, known as the coefficient omega (ω). Researchers are beginning to shift to this new indicator of internal consistency (for a practical guide to calculating omega, see Dunn, Baguley, & Brunsden, 2014).Page 95

Another form of internal consistency is split-half reliability. Just like Cronbach's alpha and the coefficient omega, split-half reliability attempts to determine the degree to which all the items in a scale are related to one another. Examining split-half reliability involves splitting the items in a scale into two parts based on some random process, then administering both halves to a group of people. After you score each half, you can calculate a correlation to see how well performance on one half is related to performance on the second half. If the measure is reliable and all items are measuring the same construct, you'd expect performance on one half of the scale to be strongly related to performance on the second half.

Inter-rater Reliability

In some research, raters observe behaviours and make ratings or judgments. For example, a researcher could ask raters to observe children playing on a playground and rate the level of compassionate behaviour demonstrated by children of different ages. Or a researcher might videotape participants getting to know one another (with their consent, of course), and then raters could code these videos by noting how often each person asks a question of the other person. To make these ratings, raters follow a strict set of guidelines in order to make these judgments as systematic as possible. One way to improve the reliability of these ratings is to have more than one person act as a rater. The reliability of these rating can then be determined by calculating *inter-rater reliability*. Inter-rater reliability is the extent to which raters agree in their observations, so if one rater gives a high score for a target (on compassion, for example), the other ratings also rate this behaviour as high. High inter-rater reliability is obtained when most of the observations for the two (or more) raters result in the same judgment. A commonly used indicator of inter-rater reliability is a statistic called *Cohen's kappa*, although alternatives certainly exist (as reviewed in the *Canadian Journal of Statistics*; Banerjee, Capozzoli, McSweeney, & Sinha, 1999). You can think of Cohen's kappa as another form of correlation that is superior to the Pearson correlation coefficient when it comes to measuring inter-rater reliability (McHugh, 2012).



LO2 Reliability and Accuracy

of Measures

Operationalizations need to be reliable, but reliability is not the only characteristic of a measure that researchers worry about. Reliability tells us whether a measure is stable or consistent in its measurement, but it does not tell

us whether our measure is measuring what it is supposed to be measuring. To use a silly but illustrative example, suppose we want to measure people's fear of heights. In order to do so, we weigh them on a scale (like at the doctor's office) and call this our "Fear of Heights Measuring Machine." Because our scale is a very good scale, it's got great reliability: It's very consistent in the measurements it gives. However, is this an accurate measure of a fear of heights? Absolutely not! The lesson here is that a measure can be highly reliable, but that doesn't mean it's measuring what it's intended to measure. Reliability is just one aspect of good measurement, but it's not the complete picture. Moreover, just because we call something a measure of a certain construct, doesn't mean it actually measures that actual construct. Our "Fear of Heights Measuring Machine" is not a *valid* indicator of intelligence.



LO3 Validity of Measures

Once shown to be reliable, the second requirement of a quality operationalization is validity. Measures are valid when they measure what they are intended to measure: They are a “true” indicator of the construct we’re interested in. We have already encountered one type of validity in [Chapter 4](#), internal validity, which reflects the degree to which an experiment is well-designed and can support a causal claim. When it comes to operationalizations, the broadest and overarching form of validity is known as [*construct validity*](#), which refers to whether a variable’s operationalization is accurate in capturing the intended phenomenon. It is the degree to which the operationalization of a variable reflects the true theoretical meaning of the variable. A measure of shyness should be an accurate indicator of that trait, and not something else or a mix of different things in addition to shyness. In terms of measurement, construct validity is a question of whether the measure that is employed actually measures the construct it is intended to measure.

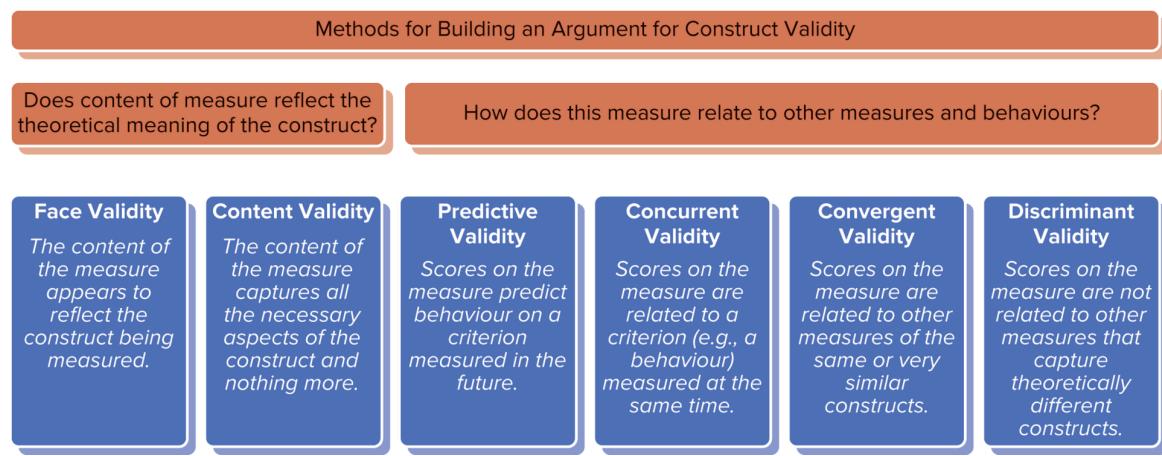
A self-report measure rating the severity of past shy behaviour is an operationalization of the construct of shyness. The validity of this measure is determined by whether it actually measures this construct in a true and meaningful way. Variables can be measured and manipulated in a variety of different ways, and there is never a perfect operationalization of a variable. Thus,

different indicators of construct validity are used to build an argument that a construct has been accurately operationalized and is properly measured by a particular scale.

Indicators of Construct Validity

How do we know that a measure is a valid indicator of a particular construct? We gather construct validity information by examining many different forms of validity. Establishing the different forms of validity help us to build an overall case for the broader category of validity, and these different forms of validity are summarized in [Figure 5.3](#). To illustrate each method, we will consider relevant research from some common measures of the construct of psychopathy. Some characteristics of psychopathy include charm, low empathy, impulsivity, and willingness to break rules and manipulate others without remorse. Some psychopaths are violent serial killers (such as Ted Bundy or Paul Bernardo), but others are not violent and can be found in everyday contexts (e.g., in the business world).

Figure 5.3 Indicators of construct validity of a measure



Face Validity

The simplest way to argue that a measure is valid is to examine whether the measure appears to assess the intended variable. This is called [*face validity*](#): The measure appears, “on the face of it,” to measure what it is supposed to measure. Face validity is not very sophisticated as it involves only a judgment of whether,

given the theoretical definition of the variable, the content of the measure appears to measure this variable. Let's consider the *Psychopathy Checklist-Revised (PCL-R)*; Hare, 1991), an inventory of traits and behaviours assessed by way of an interview (and other forms of information), with these traits including low empathy, pathological lying, impulsive behaviours, and charm. All of these appear closely related to the concept of psychopathy, suggesting this measure has high face validity. For the sake of example only, let's pretend that research finds that psychopaths prefer the colour red. In this case, an item assessing colour preference might be included in a measure of psychopathy. This item would reduce the face validity of this measure, because on the surface it doesn't make intuitive sense that colour preference has anything to do with psychopathy, even if we know it to be predictive based on research. Note that the assessment of face validity is a subjective process. But this can be improved upon by seeking out experts in the field to judge face validity. Page 97

Face validity alone is not sufficient to conclude that a measure has construct validity. Appearance is not a perfect indicator of accuracy, as something might only look like an appropriate or accurate measure. Some very poor measures may have face validity, but fall short of real construct validity. For example, most quizzes in popular magazines typically have several questions that look reasonable, but often the quiz itself doesn't tell you anything meaningful about what it's intending to measure. Interpreting the scores may make for a fun diversion, but you wouldn't learn much true about yourself from these quizzes. In addition, some good measures do not have obvious face validity. One example is that rapid eye movement (REM) sleep is an accurate (i.e., valid) indicator of dreaming, but on the face of things it is not obvious why this should be so.

Content Validity

Content validity is evaluated by comparing the content of the measure with the theoretical definition of the construct, ensuring that the measure captures all aspects of the construct and nothing extraneous to the construct. If the construct you're interested in has three different aspects, your scale should try to measure all three of these aspects, and nothing else that isn't a part of this construct. For psychopathy, it is argued that this construct involves interpersonal callousness and unemotionality, as well as impulsivity and antisocial behaviour (i.e., rule-breaking) (Guay, Ruscio, Knight, & Hare, 2007). To assess content validity in this case, we'd want to make sure that our measure of psychopathy includes questions measuring all four of those aspects—and nothing more.

Both face validity and content validity focus on assessing whether the content of a measure reflects the meaning of the construct being measured. Other aspects of validity rely on research that examines how scores on a measure relate to other measures (convergent and divergent validity) and behaviours (predictive and concurrent validity).

Predictive Validity

Another aspect of construct validity involves seeing if the measure can usefully predict some future behaviour that is theoretically related, known as *predictive validity*. This is also sometimes referred to as *criterion validity*, with the idea that the measure is predicting some important future standard, objective, or criterion. One simple example is that we would expect a self-report measure of academic motivation completed at the start of a term to predict final grades at the end of the term. These future grades that are being predicted are the standard or criterion by which we are judging the validity of our measure. Predictive validity is clearly important when studying measures designed to improve our ability to make predictions about different behaviours. Importantly, the future behaviour being predicted must be theoretically relevant to the construct of interest. For example, a study in Canada examined whether a measure of psychopathy, the PCL-R, could predict recidivism (i.e., the likelihood of committing another crime after being released from prison). Adult male Canadians who had been convicted of at least one criminal offence were tracked for ten years (Wormith, Olver, Stevenson, & Girard, 2007), and their original PCL-R scores were successful in predicting whether they were later convicted of another offence. In fact, these scores were also able to predict the length of their sentence, demonstrating that the severity of psychopathy (as measured by the PCL-R) predicts more severe criminal offences. Because future convictions are exactly what one would expect from criminals who are truly psychopathic, this criterion is a theoretically related future behaviour.

Page 98

Concurrent Validity

Concurrent validity is similar to predictive validity in that it examines prediction of a criterion, but instead of a future behaviour it examines a criterion measured at the same time as the measure is administered (i.e., concurrently).

Demonstrations of concurrent validity can take many forms. One common method is to study whether two or more groups of people differ on the measure in expected ways. For example, a self-report psychopathy scale intended for use in non-criminal populations identified more psychopathic tendencies among

university students who plagiarized than those who did not (Williams, Nathanson, & Paulhus, 2010). Because scores on this scale distinguished people who acted unethically from people who did not, this study was successful in demonstrating concurrent validity for this measure of psychopathy.

Another approach to establishing concurrent validity is to study how people who score either low or high on the measure behave in different situations. For example, you could ask people who score high versus low on a measure of psychopathy to divide 20 dollars worth of loonies between themselves and another participant (i.e., the ultimatum game). You may expect people who score low on your psychopathy measure to divide the money relatively evenly, exhibiting a sense of fairness and empathy for the other person. In contrast, you would expect psychopaths to take all of the money and not feel badly about doing so (for a similar design, see Osumi & Ohira, 2010). Finding this kind of difference in behaviour measured concurrent to our measurement using this scale lends support for the overall construct validity of the measure.

Convergent Validity

Any given measure is but one particular operationalization of the construct of interest. Often there will be other operationalizations—other measures—of the same or similar constructs. *Convergent validity* is the extent to which scores on the target measure are related to scores on other measures of the same construct or similar constructs. Different measures of similar constructs should “converge,” or be related to one another. So, for example, one measure of psychopathy should correlate highly with another psychopathy measure or measures of similar constructs. When the original PCL was found to correlate positively with antisocial personality disorder (ASPD), this was a demonstration of convergent validity (Hart & Hare, 1989). Because ASPD and psychopathy are related constructs, although distinct, we’d expect them to converge (i.e., be correlated). They share some aspects (i.e., antisocial behaviours and impulsivity), but not others (i.e., lack of remorse and empathy).

Discriminant Validity

Discriminant validity is kind of like the opposite of convergent validity, in that it is a demonstration that the measure is *not* related to variables that are conceptually *unrelated* to the construct of interest. In other words, the measure should *discriminate* between the construct being measured and other, unrelated, constructs. This form of validity has also been referred to as *divergent validity*, as

scores on the measure should diverge rather than converge with these measures of unrelated constructs. When it comes to the development of the PCL, it was important that ASPD scores were only weakly related to the part of the PCL that measures the unique, emotional aspects of psychopathy (e.g., lack of remorse and empathy) (Hart & Hare, 1989). Further discriminant validity was established in the form of weak or null correlations with other mental illnesses (e.g., schizophrenia). Although psychopathy and schizophrenia are both mental illnesses, they are so different in manifestation that we would not expect the two to converge or be related. These demonstrations of discriminant validity were important because it helped to show that the PCL was not measuring “presence of mental illness” in general, but instead was measuring psychopathy specifically.

Page 99

Collecting Evidence for Construct Validity

To build a strong argument that a measure is a valid operationalization of a construct, researchers use as many of the above methods as possible to gather as much evidence as possible. Even then, a measure may be considered well validated for use in one population or context, but that does not mean it will be a valid indicator of that construct in a different population or context. For example, a self-report scale developed to identify psychopathy in criminal populations is not necessarily appropriate to measure psychopathic tendencies among undergraduate students (Hare, Harpur, & Hemphill, 1989). Construct validity is built from various sources of evidence, and must be reconsidered whenever a measure is used for a new purpose.

☆ Student Spotlight: Scale Validity & Reliability ☆

At the University of Alberta, Jonathan Dubue, along with Dr. Chris Westbury (university supervisor) and other collaborators, wanted to develop a self-report measure of empathy to help identify students who could become peer support workers. These individuals would help other students struggling with stress and other mental health issues, and so it was important that they be high in empathy. In the process of developing their scale, they pursued evidence for both reliability and validity. For example, they investigated reliability by having the same individuals complete their measure at two different time points, three months apart. In addition, they looked at whether their questionnaire could predict scores on a different, task-based measure of empathy: the “Reading the Mind in the Eyes” task (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001). You can

read more about their scale in the *Canadian Journal of Counselling and Psychotherapy* (Dubue, Cheng, Vuong, & Westbury, 2018).



Think about It!

Read the Student Spotlight piece about scale validity and reliability. What kind of reliability was being examined in this study? What form of validity were the researchers trying to establish?



LO4 Reactivity of

Measures

One potential problem when measuring behaviour is that people can behave differently when they know they are being observed. This is known as reactivity, because people are reacting to the act of measurement and changing their behaviour. If this occurs, then we are no longer learning about how someone would really behave in the real world, only how they would behave when they know they are being observed. Consider as an example a visit to the doctor's office. For some people, having the doctor begin to measure their blood pressure by attaching that cuff to their arm can cause stress, increasing their blood pressure. The act of measuring a person's blood pressure, in this example, causes that blood pressure to change (i.e., increase). Or imagine that a researcher would like to study cheating behaviour in the context of a solitary card game. Normally you would allow yourself to cheat, but because you know you're being watched, you might alter your behaviour and play by the rules. Or a person might lie on a questionnaire to avoid feeling embarrassed, for example. Knowing that a researcher is observing you or recording your behaviour might change the way you behave, for some behaviours and for some forms

of measurement (see also [Chapter 6](#)). Measures of behaviour vary in terms of their potential reactivity. There are also ways to minimize reactivity, such as allowing time for people to become used to the presence of the observer or the recording equipment.

Another way to minimize reactivity is to measure something without that person noticing or knowing (Webb, Campbell, Schwartz, Sechrest, & Grove, 1981). These are known as *nonreactive* or *unobtrusive* operationalizations. Many such measures involve clever ways of indirectly recording a variable. For example, one way to measure preferences for paintings at an art gallery would be to ask patrons. But doing so would alert them to what you're measuring, and they might not report their preferences accurately (e.g., feeling compelled to say that they like the work of a famous artist). An unobtrusive measure of the same construct could be the frequency with which tiles around each painting must be replaced, with the most popular paintings attracting the most foot traffic and therefore incurring the most wear. Another example is a study by Levine (1990), who studied the pace of life in cities with indirect measures such as the accuracy of bank clocks and the speed of processing requests at post offices. Just like all measures, unobtrusive measures must be valid indicators of the constructs they purport to measure. Researchers must be aware of the potential problem of reactivity and reduce it whenever possible. We will return to this issue at several points in this book (e.g., [Chapters 6](#) and [9](#)).



LO5 Variables and Measurement Scales

Recall from [Chapter 4](#) that there must be at least two values (or levels) for any variable. These values may be quantitatively different (e.g., a numerical difference), or they may reflect categorical differences. However, things are actually a bit more complex, in that a variable's levels can be conceptualized in terms of four different kinds of measurement scales: nominal, ordinal, interval, and ratio (summarized in [Table 5.1](#)). You can think of these measurement scales as different kinds of variables. A variable's scale will depend on the way it is measured or manipulated (i.e., its operationalization). As we will discuss, the type(s) of scales you use will affect the rest of your research, including the conclusions you can draw, the options available for establishing construct validity, and the kinds of statistical analyses that are possible and appropriate to use when analyzing your data (see [Chapter 13](#)).

Table 5.1 Scales of Measurement

Scale	Defining Features	Example	Distinction
-------	-------------------	---------	-------------

Scale	Defining Features	Example	Distinction
Nominal	Categories with no numeric values	Canadian/Ethiopian/Lithuanian Experimental condition/control condition	Impossible to define any values and/or differences between/across categories
Ordinal	Rank ordering Numeric values have limited meaning Numeric properties have literal meaning	2-, 3-, and 4-star restaurants Birth order	Intervals between items are not known or vary in size/length
Interval	Assume equal intervals between values Zero indicates absence of variable	Intelligence Temperature (Fahrenheit or Celsius)	No true zero-point, in that zero does not indicate a complete absence
Ratio	measured Assume equal interval between values	Reaction time Age Frequencies of behaviours	Can form ratios (e.g., someone responds twice as fast as another person)

Nominal Scales

A *nominal scale* has no numerical or quantitative properties. Instead, categories or groups simply differ from one another (sometimes nominal variables are called *categorical variables*). An example is the variable of country of birth. People are born in a certain country, and we can classify people based on what country they were born in. However, these categories don't have numerical properties, and you can't be more or less than when it comes to country of birth: The levels are merely different. Another example is the classification of undergraduates according to major. A psychology major would not be entitled to a higher number than a history major, for instance. Even if you were to assign numbers to the different categories, the numbers would be meaningless, except for identification. This is called a nominal scale because nominal means "existing in name only," and we assign names to different categories within a variable.

In an experiment, the independent variable is often a nominal or categorical variable. For example, one study examined whether different task goals affected learning (Cianci, Klein, & Seijts, 2010). Students about to take an analogy test were given either a performance goal (i.e., earn the highest score) or a learning goal (i.e., use the test to learn how to answer these types of questions). In this study, the condition variable is nominal because the two levels are merely different categories: The two goals have no numerical properties. After everyone was told they failed on the first attempt, those who had the learning goal in mind performed better after failure than did those with the performance goal.

Ordinal Scales

An *ordinal scale* allows us to order the levels of the variable in terms of rank. Instead of having categories that are simply different, as in a nominal scale, the categories can be ordered from first to last. One example of an ordinal scale is Olympic medals (gold, silver, and bronze) and another is birth order (first, second, third, and so on). The developmental stages

proposed by Piaget (1952) are also ordinal, theorizing that intellectual development progresses sequentially through four different stages.

Importantly, although we can rank ordinal elements, we don't know anything about the distance or difference between each element. With birth order, although we know that the first-born is older than the second-born, we don't know how much older (unless we ask, which means getting another kind of information). We cannot say that the difference in age between a first-born and a second-born is always the same, or that it is the same as the difference between a second-born and a third-born. As another example, in health psychology, researchers might distinguish among high, medium, and low socio-economic status (SES). No particular value is attached to the intervals between the numbers used in ordinal scales.

Interval Scales

In an *interval scale*, the difference between the numbers on the scale are equal in size. The difference between 1 and 2, for example, is the same size as the difference between 2 and 3. Interval scales generally have five or more quantitative levels. For example, a household thermometer measures temperature on an interval scale. The difference in temperature between 15°C and 17°C is equal to the difference between 26°C and 28°C. Another feature of interval scales is that zero does not indicate a complete absence of quantity. Taking the example of a thermometer once again, there is no absolute zero on the scale that would indicate the absence of any temperature. For interval scales, the zero is only an arbitrary reference point. In fact, 0° Celsius equals 32° Fahrenheit! One consequence of this is that we cannot form ratios based on these numbers; that is, we cannot say that one number on the scale represents twice as much (or three times as much, and so forth) temperature as another number. It would not be correct to state that 20°C is twice as warm as 10°C, for example. This point becomes even more clear when these very same values are converted to Fahrenheit: 68°F (20°C) is clearly not twice as warm as 50°F (10°C).

Page 102

Scores on the Weschler intelligence test can be considered interval scales. The difference between an IQ score of 85 and 90 is the same as a difference

between 105 and 110—but there is no absolute zero point that indicates an absence of all intelligence. Moreover, ratings scales (e.g., 7 points ranging from 1 to 7; see [Chapter 7](#)) are often assumed to be interval scales (or to behave as interval scales, when averaging across many responses). These kinds of scales are often used to measure personality traits such as agreeableness, or attitudes toward different groups of people. If the measurement is an interval scale, we cannot make a statement such as “the person who scored 6 is twice as extraverted as the person who scored 3” because there is no absolute zero point that indicates an absence of the trait being measured.

Ratio Scales

A *ratio scale* is like an interval scale, except it does have a meaningful absolute zero point that indicates total absence of the variable being measured. One example can be observed with respect to temperature and the Kelvin scale, in which the “absence of temperature” is defined as the point at which all molecular movement stops: 0° Kelvin (-273° Celsius). Many physical measures also use ratio scales, such as length, weight, or time. Ratio scales enable statements such as “a person who weighs 100 kilograms weighs twice as much as a person who weighs 50 kilograms” or “on average, participants in the experimental condition responded twice as fast those the control condition.” These types of comparisons in the form of ratios are not possible when using other scales.

Ratio scales are often used in the behavioural sciences when variables that involve physical measures are being studied. This would include measures of time (e.g., duration and reaction time), rate of responding, and heart rate. However, many variables in the behavioural sciences cannot be measured with such precision and so use nominal, ordinal, or interval scale measures.

The Importance of the Measurement Scales

With this information about measurement scales, when you now read about measures and operationalizations, you should try to identify the type of measurement scale. As we will consider in [Chapters 12](#) and [13](#), the

measurement scale determines the types of statistics that are appropriate to use when analyzing the results of a study. Moreover, the conclusions one draws about the meaning of a particular score on a variable depend on which type of scale was used. With interval and ratio scales, you can make quantitative distinctions that allow you to talk about amounts of the variable. With nominal scales, there is no quantitative information. To illustrate, suppose you are studying perceptions of physical attractiveness. In an experiment, you might show participants pictures of people with different characteristics (e.g., different eye colour and hair colour combinations). How should you measure these judgments of physical attractiveness? One possibility is that you could use a nominal scale:

- _____ Not Attractive _____ Attractive

Page 103 These scale values allow participants to state whether they find the person attractive, but do not tell you anything about how attractive they find the person. As an alternative, you could use an interval scale that asks participants to rate the amount of attractiveness:

- Very Unattractive _____ Very Attractive

This rating scale provides you with quantitative information about the amount of attractiveness because you can assign relatively meaningful numeric values to each of the response options on the scale.

We are now ready to consider approaches to operationalizing behaviour. A variety of observational methods are described in [Chapter 6](#), with questionnaires and interviews the focus of [Chapter 7](#).



Illustrative Article: Measurement Concepts

Every term, millions of students complete course evaluations in an effort to assess the quality and performance of their instructors. This specific measurement instrument can vary from campus to campus, but the overall goal is the same. Course evaluations are used to inform hiring decisions,

promotion decisions, and classroom instruction decisions, and they are also used by individual instructors to improve the courses that they teach.

Brown (2008) was interested in student perceptions of course evaluations. He collected data from 80 undergraduates enrolled in an undergraduate research methods course and examined their perceptions of student evaluations of teaching, of mid-semester evaluations, and of the effectiveness of completing mid-semester evaluations.

He found, among other things, that although participants believed that students are honest in their evaluations and that the evaluations are important in hiring decisions, they were less sure that instructors took the evaluations seriously and also tended to believe that students evaluate courses based on the grade that they get, or to “get back” at instructors.

For this exercise, acquire and read the following article:

- Brown, M. (2008). Student perceptions of teaching evaluations. *Journal of Instructional Psychology*, 35(2), 177–181.

After reading the article, consider the following:

1. Brown did not report any reliability data for his measures. How would you suggest that he go about assessing the reliability of his measures?
2. In the context of evaluating college and university teaching, how would you describe the construct validity of course evaluation measures generally (or the specific tool that is used on your campus)? That is, how well do student evaluations truly assess the construct of course quality? Specifically, how would you assess the content, predictive, concurrent, convergent, and discriminant validity of student course evaluation measures?
3. Brown did not report any validity information for his measures of participant perceptions. Assess the face validity of his measures.
4. Do you think that Brown’s measures are reactive? How so? Likewise, do you think that course evaluations are reactive? How so?

5. Describe the level of measurement used in Brown's study. Generate two alternative strategies for measurement that would occur at different levels.

Study Terms

Test yourself! Define and generate an example of each of these key terms.

- *coefficient omega* (p. 94).
- *concurrent validity* (p. 98).
- *construct validity* (p. 96).
- *content validity* (p. 97).
- *convergent validity* (p. 98).
- *Cronbach's alpha* (p. 94).
- *discriminant validity* (p. 98).
- *face validity* (p. 96).
- *internal consistency reliability* (p. 94).
- *inter-rater reliability* (p. 95).
- *interval scale* (p. 101).
- *measurement error* (p. 91).
- *nominal scale* (p. 101).
- *ordinal scale* (p. 101).
- *Pearson correlation coefficient* (p. 93).
- *predictive validity* (p. 97).

- *ratio scale* (p. 102).
- *reactivity* (p. 99).
- *reliability* (p. 90).
- *split-half reliability* (p. 95).
- *test-retest reliability* (p. 93)
- *true score* (p. 91)

Review Questions

Test yourself on this chapter's learning objectives. Can you answer each of these questions?

1. How is the quality of an operationalization affected by low reliability and low validity?
2. What is meant by the reliability of a measure? Distinguish between true score and measurement error.
3. Compare and contrast the three ways to determine the reliability of a measure. When would a researcher use each kind of reliability?
4. Discuss the concept of construct validity, including how a researcher builds an argument for it.
5. Compare and contrast convergent and discriminant validity, and predictive and concurrent validity.
6. Why isn't face validity sufficient to establish the validity of a measure?
7. What is a reactive measure?
8. Distinguish among nominal, ordinal, interval, and ratio scales.

Deepen Your Understanding

Develop your mastery of these concepts by considering these application questions. Compare your responses with those from other people in your study group.

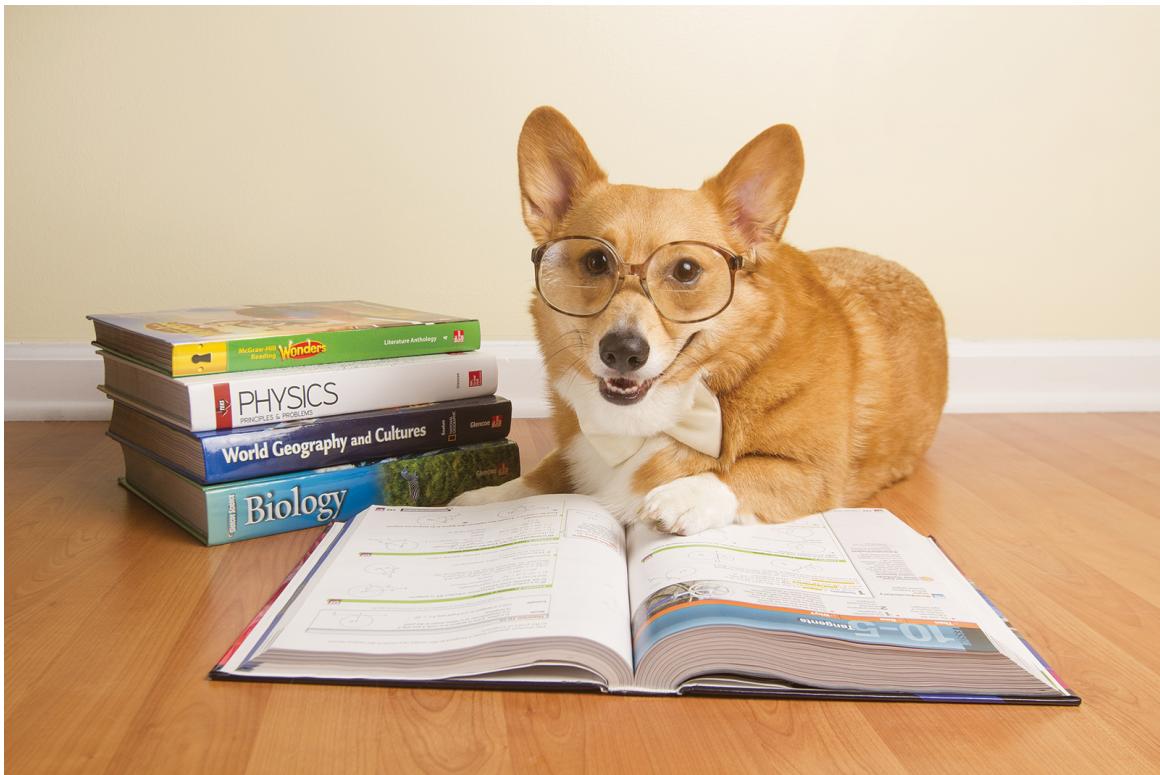
1. Find a reference book on psychological measurement such as the one by Robinson, Shaver, and Wrightsman (1991), or do a *PsycINFO* search (see [Appendix D](#)) using a keyword search for “scale” AND “validity.” Identify a measure that interests you and describe the ways the researchers built a case for its reliability and validity.
2. Here are a number of references to variables. For each, identify whether a nominal, ordinal, interval, or ratio scale is being used:
 1. Today’s temperatures in the eight biggest cities across Canada.
 2. The birthweights of babies who were born at Whitehorse General Hospital last week.Page 105
 3. The number of hours you spent studying each day during the past week.
 4. The amount of tip left after each meal served at a restaurant during a three-hour period.
 5. The number of votes received by each candidate for provincial parliament in your riding in the last election.
 6. The brand listed third in a consumer magazine’s ranking of tablet computers.
 7. In a sportswriter opinion poll, the Montreal Canadiens were listed as the number one Canadian hockey team, with the Calgary Flames listed number two.

8. Your friend's score on an intelligence test.
 9. Yellow walls in your office and white walls in your boss's office.
 10. The type of programming on each radio station in your city (e.g., in Winnipeg, CITI plays rock, CBW is talk radio).
 11. Ethnic group categories of people in a neighbourhood.
3. Take a personality test on the Internet. Based on the information provided, what can you conclude about reliability, construct validity, and reactivity?
4. Think of an important characteristic that you would look for in a potential romantic partner, such as humour, intelligence, attractiveness, hard work ethic, religiousness, and so on. How might you measure that characteristic? Describe two methods that you might use to assess construct validity of that measure.

Chapter 6

Page 106

Observational Methods



©McGraw-Hill Education/Holly Hildreth

Science is the process of systematic observation, but there are many factors to consider when observing. If you want to study students, for example, you might want to consider posing as a student yourself, just like this dog.

Learning Objectives

Keep these learning objectives in mind as you read to help you identify the most critical information in this chapter.

By the end of this chapter, you should be able to:

1. [LO1](#) Compare quantitative and qualitative approaches to investigating behaviour.
2. [LO2](#) Describe naturalistic observation as a method and discuss related issues such as participation and concealment.
3. [LO3](#) Describe systematic observation as a method and discuss related issues such as the use of coding schemes, participant reactivity, equipment, reliability, and sampling.
4. [LO4](#) Describe the features of a case study and its appropriate uses.
5. [LO5](#) Describe archival research and various sources of archival data.

Page 107Have you ever “people-watched”? This is the act of sitting back and simply observing those around us, and it can be fascinating. Just think about the wide variety of behaviours you have observed in university dormitories, on public transit, at the grocery store, on the street, on social media, at concerts or sporting events, and so on. In this chapter, we will explore various techniques that behavioural scientists have developed to turn everyday observations into knowledge. These observation techniques are used to generate hypotheses for further research and to provide important in-depth descriptions of phenomena ([Chapter 1](#)), both in non-experimental research (in which behaviour is measured or observed rather than manipulated) and when measuring outcomes in experimental designs (e.g., measuring the dependent variable; [Chapter 4](#)).

Many of the methods described in this chapter are used in different ways in many different fields. For example, you may have encountered some of these methods in your anthropology or sociology courses. Therefore, it will be helpful to start by situating these methods in the context of quantitative and qualitative perspectives.



LO1 Quantitative and Qualitative Approaches

The goal of any researcher is to better understand some real-world phenomenon. This is typically achieved by virtue of systematic observation of some kind. However, there are two broad approaches to making these observations, based on the type of data collected: quantitative approaches and qualitative approaches. Quantitative approaches involve collecting data in the form of numbers (i.e., numerical data). Much of what we've been discussing so far maps rather well onto quantitative approaches, with self-report surveys, reaction-time tasks, and psychophysiological measures (e.g., heart rate) all resulting in numerical data (e.g., scores on a response scale, reaction times in milliseconds, and heart rate in beats per minute). These data are then analyzed using statistics. This quantitative approach may be most familiar to you, but it is not the only approach.

Another way to learn about the world is to collect data in non-numerical form, using an *interpretive* (Poulin, 2007) or qualitative approach. A qualitative approach offers a deep description of people's behaviour in natural settings, through people explaining their experiences in their own words. Often, this approach involves collecting in-depth information on relatively few individuals or within a limited setting, and conclusions are

based on careful interpretations drawn by the investigator. Relative to quantitative approaches, there is a greater variety in the paradigms and procedures associated with qualitative approaches (e.g., ethnography, phenomenology, grounded theory, narrative; Creswell, 2007).

Imagine that you are interested in describing how the lives of teenagers are affected by working outside of school. Taking a quantitative approach might involve developing a questionnaire and asking a sample of teenagers to complete it ([Chapter 7](#)). You could ask about the number of hours they work, the type of work they do, their stress levels, and their school grades. Using the numerical scores for this survey, you would analyze the data using statistics. These analyses would produce a report of the results, which might include things like average levels of stress, and the relationship between hours worked and academic average as described by a correlation coefficient.

Page 108

Suppose, instead, that you decide to take a qualitative approach to understanding this same phenomenon. In this case you might organize a series of focus groups in which you interview groups of teenagers about their experiences working outside of school while also being a student. The teenagers participating would then tell you about their experiences using their own words. You might even record the discussions and have their words transcribed into a text document for further analysis later on, or have observers take detailed notes during the discussions. Because the data being collected are not numbers, but instead the actual words of teenagers themselves, this is a qualitative approach. In fact, the collection and interpretation of any form of data other than numbers constitutes a qualitative approach (e.g., photographs on an Instagram account, audio recordings, drawings). The results of this qualitative study would look somewhat different from a quantitative approach. Instead of a statistical analysis of numbers, a qualitative report may describe different themes that emerged from the discussions, for example.

Note that in some cases, qualitative data might be collected (e.g., transcribed interviews) and then converted into numbers for a quantitative statistical analysis (e.g., frequency of using personal pronouns like “I”). We could consider this a type of qualitative approach since non-numerical data

are being collected, although the main thing to recognize is that we’re beginning with qualitative data (i.e., non-numerical information), then transforming it into numbers for a quantitative analysis. In addition, other methods, both qualitative and quantitative, could also be used to study teenage employment. For example, a quantitative study could examine archival data collected from Statistics Canada, or a qualitative study might use concealed naturalistic observation in the form of researchers working as a trainee in a fast-food restaurant.

As we explore various methods in this chapter and in [Chapter 7](#), consider how each may be used from quantitative and qualitative perspectives. Some of the methods are used more commonly by investigators using a qualitative approach (e.g., naturalistic observation, focus groups), whereas some are used more commonly by investigators using a quantitative approach (e.g., systematic observation, surveys collecting numerical responses). Many of these methods are used for both qualitative and quantitative purposes, but just in different ways (e.g., case studies, archives). Keep in mind that what counts as “data” is one of the keys to distinguishing the two approaches: The quantitative approach requires statistical analysis and hence numerical data, whereas the qualitative approach often involves interpreting people’s experiences and so can be applied to non-numerical data. Whether a researcher uses quantitative or qualitative approaches to learning about the world—or uses both (known as a *mixed-method approach*)—depends on many factors, including the kind of training one receives as a student. Because it often takes years of training to become an expert in using either approach to research, qualitative and quantitative researchers sometimes fail to appreciate the strengths of the other approach. In the big picture, however, a thorough understanding of real-world behaviour will likely require both qualitative and quantitative ways of knowing. Moreover, some questions might be better addressed by one approach or another. For example, if we’re interested in how often something occurs, or how much of something exists, then quantitative approaches would be an obvious choice as these quantities translate easily into numbers (e.g., two shifts a week, \$15 per hour). In contrast, if we’re more interested in how people feel about something, or why they do something, then qualitative approaches might be a good way to capture the complexity of these thoughts, feelings, and experiences.



LO2 Naturalistic

Observation

Naturalistic observation is a research method in which researchers immerse themselves in a particular natural setting (i.e., the real world, or what is sometimes called *the field*). Observations are typically made over an extended period of time using a variety of information-collection techniques. Quantitative data can be collected in the form of numerical information (e.g., how many patrons of an ice cream parlour are wearing shorts), or qualitative data can be gathered in the form of detailed field notes. These field notes can include information about all aspects of a situation, including the setting, patterns of personal relationships, people's reactions to events, and so on. This qualitative data-gathering need not rely solely on passive observation; it can also involve interviewing key "informants" to gain inside information, talking to people about their lives, taking pictures of the environment (e.g., the types of posters placed in a break room), and examining documents produced in the setting, such as newsletters, e-mails, or manuals. Qualitative field researchers can even collect audio and video recordings to gain a holistic picture of a situation and the people within it. Naturalistic observation has roots in anthropology

and the study of animal behaviour, and is used in the social sciences to study many phenomena in all types of social and organizational settings.

The first goal of researchers using this technique is to describe the setting, events, and persons observed. The second, equally important, goal is to interpret what was observed. Depending on the specific qualitative approach used, this interpretation may involve identifying common themes (as in phenomenology) or developing a theory that can generate hypotheses for future work (as in grounded theory; see Creswell, Hanson, Clark, & Morales, 2007). The final report might reflect the chronological order of events (as in the narrative approach), or it can be organized around the theory developed by the researcher (as in grounded theory). Specific examples of events that occurred during observation are often used to support and illustrate the researcher's interpretations. A good naturalistic observation report will support the validity (i.e., accuracy) of the interpretation by using multiple sources of confirming evidence (Creswell, 2007). For example, similar events may occur several times, similar information may be reported by two or more people, and several different events may occur that all support the same interpretation or conclusion. The published qualitative report includes both specific observations and the researcher's interpretation.

Researchers can use naturalistic observation to describe and understand how people in a social or cultural setting live, work, and experience this setting. For example, to learn about the hotel community, qualitative researchers have spent months visiting hotels, talking to people, observing interactions, and becoming accepted as "regulars" (Prus & Irini, 1980). In another particularly illuminating example of naturalistic observation, eight psychologically healthy people were admitted to psychiatric hospitals after claiming to "hear voices" but exhibiting no other symptoms, in order to learn about the conditions and treatment of mental health patients (Rosenhan, 1973). This study revealed systematic problems with diagnosis and treatment of mental illness, including deep and long-lasting stigmatization.

☆ Student Spotlight: Naturalistic Observation ☆

While an undergraduate student at Simon Fraser University, Natalie Harrison became interested in how individuals seeking treatment for mental health issues perceive their reception and care at hospital emergency departments. In collaboration with Dr. Ronald Roesch and others, they conducted interviews with almost 50 patients who had visited the emergency department of a hospital in British Columbia for psychological or psychiatric distress. Their main goal was to learn how patients perceived their experience, in order to provide guidance on how emergency departments might modify their procedures to better accommodate these individuals. The results of this in-depth field observation of real patients now appears in the *International Journal of Forensic Mental Health* (Harrison, Mordell, Roesch, & Watt, 2015).

Naturalistic observation is frequently employed by researchers using a qualitative approach, as noted above. In these cases, the data are primarily qualitative descriptions based on systematic observations. Such descriptions can be richer, more detailed, and closer to the phenomenon being studied than statistical representations of quantitative data. Sometimes researchers use a mixed approach, gathering some quantitative data in addition to their qualitative data. If circumstances allow, for example, data can be gathered on income, family size, education levels, and other easily quantifiable variables for the persons under observation. A researcher using a mixed approach might report and interpret these quantitative data along with the qualitative data gathered from interviews, photographs, and so forth.

Page 110

Field researchers using naturalistic observation can also employ a fully quantitative approach, collecting only numerical data. In addition, naturalistic observation can often be used to generate hypotheses for later laboratory experiments. For example, to learn about how people persuade or influence others, psychologist Robert Cialdini was hired and trained in many sales positions (e.g., selling cars), experiences that were very helpful for developing his theories that he later tested using experiments (see Cialdini, 2008). Naturalistic observation can also be used to collect qualitative data that is later translated into numerical data using a *coding scheme* (described [below](#)).



Think about It!

Imagine that you would like to show someone else how smart you are. One option is to provide evidence in the form of quantitative data (e.g., numerical information such as your high school grades). Another option would be to provide qualitative data (e.g., an essay you wrote, testimony from past teachers). If you could choose only one, which would you choose and why? Would combining both be better than choosing only one? Why or why not?

Issues in Naturalistic Observation

Participation and Concealment

In naturalistic observation, there are two key issues that are also highly related to one another. The first is whether to actively participate in the situation you are studying (e.g., volunteer as a firefighter to study this community), or not participate but still observe. The second issue is whether to conceal your purpose or presence from the other people in the setting. Do you tell the rest of the group that you're a researcher interested in studying them? Do you allow people to be aware that you're present and observing?

A non-participant observer is an outsider who does not become an active part of the setting being observed. In contrast, a participant observer assumes an active, insider role. Because *participant observation* allows the researcher to observe the setting from the inside, this allows the researcher to experience events in the same way as the other participants. Participant observation may facilitate friendships and other insider experiences that offer valuable data. For example, Nathan (2005) studied the undergraduate experience by enrolling as a mature student, and she describes many insights she gained from the friendships she made while living in a dormitory. A potential problem with participant observation, however, is that the observer may lose the objectivity necessary to conduct scientific observation. Remaining objective may be especially difficult when the

researcher already belongs to the group being studied and identifies with that community. If a researcher has some prior reason to either find fault with the community being studied or to report only positive behaviours of that group, there is great risk that the observations will be biased and the conclusions will lack objectivity (see Creswell & Miller, 2000).

The second key issue is whether the researcher should be out in the open or keep his or her presence concealed. *Concealed observation* may be preferable because the presence of the observer may influence the behaviour of those being observed. Imagine how the presence of an observer might change the behaviour of a classroom of high school students. Concealed observation results in less participant reactivity compared to non-concealed observation ([Chapter 5](#)), because if people are unaware that they are being observed, they can't react to this act of observation. That said, in some cases non-concealed observation may be preferable from an ethical viewpoint. For example, consider the ethical principles that were stretched when researchers perched in a bathroom stall to record handwashing practices among university women (Drankiewicz & Dundes, 2003).Page 111

Non-concealed observation can also become non-problematic as reactivity can subside quickly. People can quickly become used to the observer and behave naturally in the observer's presence. This is particularly true when the observer is not physically present, such as when technology is used to observe and record. In some studies, researchers have asked participants to wear an audio-recording device that intermittently records snippets of ambient sound, capturing small clips of people's daily lives (Mehl, Pennebaker, Crow, Dabbs, & Price, 2001). These recordings provide a rich source of qualitative data based on naturalistic observation (Vazire & Mehl, 2008). Studies such as these have revealed that people seem to quickly forget about the recording device and spontaneously reveal many private aspects of their lives while wearing it.

The decision of whether to conceal one's purpose and/or presence depends both on ethical concerns and on the nature of the particular group being studied. Sometimes a participant observer is revealed to certain members of the group, who give the researcher permission to be part of the group as a

concealed observer. Often a concealed observer says nothing about his or her purposes spontaneously, but will completely disclose the goals of the research if asked by anyone. Non-participant observers are also not concealed when they gain permission to “hang out” in a setting or use interview techniques to gather information. In actuality, then, there are degrees of participation and concealment: A non-participant observer may not become a member of the group, for example, but may over time become accepted as a friend or simply part of the ongoing activities of the group. Researchers who use naturalistic observation to study behaviour must carefully determine what their role in the setting will be.

You may be wondering about the role of informed consent in naturalistic observation. Recall from [Chapter 3](#) that observation in public places when anonymity is not threatened and behaviour is not manipulated is typically considered exempt from ethical review. In these cases, informed consent is usually unnecessary. Nevertheless, researchers must be sensitive to ethical issues when conducting naturalistic observation. Of particular interest is whether the observations are made in a public place where there is no clear expectation that behaviours will be private. For example, should a bar be considered public or private? What about a person’s Facebook page? Participation on many Internet forums is public, but people who post may perceive that they are part of a private community, which poses tricky ethical issues (Kraut et al., 2004). Issues of informed consent also resurface when collecting data with the audio-recording device mentioned earlier, precisely because participants become so used to it. To ensure ongoing informed consent, sometimes participants are given all audio files before data analysis, so they can delete any they are uncomfortable sharing with researchers (Holleran, Whitehead, Schmader, & Mehl, 2011). There are rarely black-and-white answers for many of the questions that arise during research, including ethical questions. It is always important to evaluate each study and each issue individually, carefully considering all the relevant factors.

Defining the Scope of the Observation

A researcher employing naturalistic observation may want to study everything about a setting. However, this may not be possible simply

because settings are so complex. Thus, researchers must often limit the scope of their observations to behaviours that are relevant to the central issues of the study. For example, a naturalistic observation study of skateboarders might focus on their intrinsic motivation (Seifert & Hedderson, 2009), while ignoring other aspects of their behaviour such as their use of public spaces, their social interactions, or the experience of different genders. For another example, when Graham and colleagues (2006) used naturalistic observation to study behaviour in bars, they specifically focused on expressions of aggression, rather than attempting to observe factors related to alcohol consumption, expressions of sexuality, and friendship.

Page 112

Limits of Naturalistic Observation

Naturalistic observation obviously cannot be used to study all issues or phenomena. From the perspective of qualitative research, this approach is most useful when investigating social settings as it helps us to capture the rich complexity of these contexts. For quantitative researchers, naturalistic observation is useful for gathering data in real-life settings and generating hypotheses for later laboratory experiments. Although we cannot control the setting in naturalistic observation, making it challenging to test well-defined hypotheses under precisely specified conditions, these observations take place in the real world and have good ecological validity.

Naturalistic observation in the field is also very difficult to conduct (cf., Green & Wallaf, 1981). Unlike a typical laboratory experiment, data collection in the field cannot always be scheduled at a convenient time and place. Consider the naturalistic observation study of aggression in bars (Graham et al., 2013), in which research assistants underwent extensive training and then often stayed in and around bars until 3:30 a.m., on over 1,000 different occasions! In fact, field research in general (whether it is naturalistic observation or a field experiment; [Chapter 4](#)) can be extremely time-consuming, often placing the researcher in an unfamiliar setting for extended periods of time. Also, in laboratory research, the procedures are well-defined and identical for each participant, and with the data analysis often planned in advance. In naturalistic observation research, however, there is an ever-changing pattern of events, some important and some

unimportant. The researcher must record them all and remain flexible in order to adjust to them as research progresses. Finally, the process of interpreting these data is not simple. The researcher must repeatedly sort through vast amounts of data in order to identify common themes, develop hypotheses to explain the data, or code the data into meaningful categories for statistical analysis, depending on the approach being used (e.g., phenomenology, grounded theory, quantitative). Although naturalistic observation research is a difficult and challenging technique, it can yield valuable knowledge when done well (e.g., Rosenhan, 1973; Cialdini, 2008).



LO3 Systematic Observation

Systematic observation refers to the careful observation of one or more specific behaviours in a particular setting. However, unlike naturalistic observation, which takes place in the real world, for systematic observation, this setting is often created by the researcher. This research method is employed in far fewer disciplines than naturalistic observation, and is used more often within a quantitative rather than qualitative approach. For systematic observation, the researcher is interested in only a few very specific behaviours, the observation of these behaviours is easily quantifiable, and the researcher typically has strong hypotheses about the behaviours. Page 113

For example, a team of researchers at Dalhousie University in Halifax were interested in how young children develop empathy for others who are expressing pain or sadness (Bandstra, Chambers, McGrath, & Moore, 2011). Children aged 18 to 36 months were videotaped in a room while they played with an experimenter who either pretended to hurt herself and expressed pain, or broke a toy and expressed sadness. In this experimental design, the experimenter's emotion (pain versus sadness) was the independent variable. After the study was over, observers viewed the

videotapes and coded each child's behaviour during the 30-second period in which the experimenter expressed the emotion. Observers used the *coding scheme* depicted in [Table 6.1](#).

Table 6.1 Coding Scheme for Children's Empathic Responses
(Bandstra et al., 2011)

Code	Definition	Frequency
<i>Prosocial acts</i>	Child attempts to comfort the experimenter through distraction, by sharing a new toy, or with verbal sympathy.	
<i>Attempts to understand the distress</i>	Child imitates the experimenter's emotional expression.	
<i>Self-distress</i>	Child engages in self-soothing behaviour (e.g., sucks thumb, seeks parental comforting).	
<i>Unresponsive/inappropriate affect</i>	Child shows little concern (e.g., ignores experimenter, plays, laughs).	
<i>Global concern</i>	Overall level of concern (0 = <i>no concern evident</i> to 4 = <i>variety of responses clearly indicating concern</i>).	

This coding of behaviours was the researchers' operationalization of the dependent variable: empathic responses from the children. What they found was opposite to what had been found among adults in a previous study:

Children seemed more affected by witnessing another's sadness relative to witnessing pain. These findings suggest that children may develop empathy for different emotions at different rates.

Coding Schemes

In systematic observation, the researchers must decide which behaviours are of interest, choose or create a specific setting in which the behaviours can be observed, and develop a coding scheme to record and categorize observations, such as the one in [Table 6.1](#). In general, the purpose of a coding scheme is to quantify qualitative observations, be they live actions, written responses, or images produced by participants. Sometimes the researcher develops the coding scheme to fit the needs of the particular study (Boyatzis, 1998). Coding schemes should be as simple as possible, allowing observers to easily categorize behaviours (e.g., presence or absence of thumb-sucking). This need for simplicity is especially important when observers are coding live behaviours rather than viewing videotapes, as live actions can't be rewound and reviewed. This is why videotaping participants is often preferable, so that observers can carefully review their coding. As another example, Holleran and colleagues (2011) used the discreet audio-recording device mentioned earlier to examine job dissatisfaction among female faculty in STEM disciplines (i.e., science, technology, engineering, and mathematics). To facilitate the transformation of these qualitative data (i.e., audio recordings) into quantitative data (i.e., numbers), a very simple and precise coding scheme was employed: Conversations were coded for (1) topic (social or research-related) and (2) gender of the person to whom they were speaking. This study found that faculty were less likely to discuss research with female colleagues than with male colleagues.
Page 114

Developing a simple and precise coding scheme can be a surprisingly complex process. If you are going to code only for a few things, you want to be confident that you're coding for the right things and not missing anything important. Researchers might first choose to conduct a naturalistic observation study to help them come up with meaningful categories for their coding scheme, which they could then use to code behaviours observed in a more controlled systematic observation study. This was the

strategy pursued by Graham and colleagues (2006). After observing people in 118 Toronto bars for 1,334 nights, the researchers created a coding scheme identifying many forms of physical aggression (e.g., pushing, restraining) and non-physical aggression (e.g., swearing, glaring). This coding scheme was later used in quantitative research investigating people's motives for different types of aggressive barroom acts (e.g., Graham et al., 2013).

Sometimes researchers use coding schemes that have been developed by others. For example, the Facial Action Coding System (FACS; see Ekman & Friesen, 1978) is a way to categorize subtle, fleeting, facial muscle movements. One way that the FACS has been used in psychology is to code emotional expressions, including surprise and shame (Tracy, Robins, & Schriber, 2009). Another coding scheme is the Mealtime Interaction Coding System (MICS; Dickstein, Hayden, Schiller, Seifer, & San Antonio, 1994). The MICS provides a way of coding the interactions of family members during mealtimes, including standards of communication and behaviour. It has been used to explore how family interactions during mealtime relate to childhood attachment (Dubois-Comtois & Moss, 2008) and to children's (over-)eating behaviour (Czaja, Hartmann, Rief, & Hilbert, 2011). A major advantage of using a previously developed coding scheme is that research already exists that validates the coding scheme's use for particular purposes, and training materials are usually available.

Issues in Systematic Observation

Inter-rater Reliability

Recall from [Chapter 5](#) that *reliability* refers to the degree to which a measurement is stable, consistent, and precise. When conducting systematic observation, two or more raters are tasked with coding the behaviours. Reliability for this coding is indicated by high agreement among the raters: Both raters code the same behaviours in the same way. This is known as *inter-rater reliability*, also discussed in [Chapter 5](#). Very high levels of agreement are reported in virtually all published research using systematic observation (generally 80 percent agreement or higher; e.g., Bandstra et al., 2011). This is because coders are often trained

extensively on the coding scheme before they begin coding the data in question. This means practising on prior or example data and discussing coding decisions before coding new data (see Bakeman & Gottman, 1986).

That said, it can be difficult to achieve high inter-rater reliability with live coding, unless the coding scheme is very simple (e.g., counting the number of times a participant coughs during an interview). Video-recording behaviour has the advantage of providing a permanent record that can be coded later by many observers. It also means that coding can be double-checked by other researchers. Depending on the data and the coding scheme, computer software can also be used to count and time observed behaviours.

Participant Reactivity

Just as in naturalistic observation, the presence of the observer (be it a video camera or a live person) can affect people's behaviours by producing participant reactivity ([Chapter 5](#)). As with naturalistic observation, reactivity can be reduced by concealed observation. For example, one-way mirrors and hidden cameras can be used to conceal the presence of an observer in a laboratory testing room. However, gaining participants' informed consent to use their data is crucial after these types of covert recording ([Chapter 3](#)). It is especially important to fully debrief participants and offer them the opportunity to forbid researchers to use their data when concealed observation tools are used. ([Chapter 9](#) offers more details on debriefing.) Alternatively, reactivity can be reduced by allowing enough time for people to become used to the presence of the observer and any recording equipment.

Sampling Behaviours

Researchers must make decisions about how to sample behaviours. For many research questions, samples of behaviour taken over a long period provide more accurate and more useful data than single, short observations. Consider a study on parents' behaviour at youth hockey games that used systematic observation in a field setting (Bowker et al., 2009). The researchers wanted to know whether parents promoted violence at hockey

games, as is sometimes reported in the media. They could have studied a single hockey game or two, but such data might be distorted by short-term trends (e.g., who is playing, time of the game, whether a particularly vocal parent is present). A better method of addressing the question is to observe spectator behaviour over time, which is exactly what the researchers did. Observers attended 69 youth hockey games (approximately 10 percent of all games played by 11- to 14-year-olds in Ottawa, between November 2006 and February 2007), and recorded all comments made by spectators for each team. Then, researchers systematically coded all comments into one of five categories (e.g., positive and directed specifically at one player, instructions for plays). They found that comments tended to be positive, especially at recreational (rather than competitive) games. The general positivity of comments collected in this systematic way offers evidence against the popular belief that parents frequently, or predominantly, fuel violence in youth hockey.



LO4 Case Studies

A *case study* provides a detailed description of an individual person or a setting, such as a business, school, or neighbourhood. A case study can be based on naturalistic observation, but this is not necessarily the case. For example, a case study may be a description of a patient by a clinical psychologist or even a historical account of an event, such as an intervention program that failed. One particular form of case study is the *psychobiography*, in which a researcher applies psychological theory to explain the life of an individual, usually an important historical figure (cf., Elms, 1994; e.g., Runyan, 2006). Case studies can also involve library research and interviews with people familiar with the target person or setting, and may or may not include direct observation (cf., Yin, 1994). Sometimes direct observation of the person in question is not even possible —if the person is deceased, for example.

Depending on the goals of the investigation, a case study may present the individual's history, symptoms, characteristic behaviours, reactions to situations, or responses to treatment. The case study method is very useful when an individual possesses a particularly rare, unusual, or noteworthy condition. In these cases, studying more than one person might not even be possible or feasible. One famous case study within the study of memory involved a man with an amazing ability to recall information, as described

by Luria (1968). The man in these reports, referred to simply as “S.” could remember long lists and passages with ease, apparently using elaborate mental imagery. Luria also described some of the drawbacks of S.’s ability. For example, S. frequently had difficulty concentrating because mental images would spontaneously appear and interfere with his thinking.

Page 116

A more recent example of a case study investigated language development in deaf children, which is typically slower than among hearing children. York University researchers Ruggirello and Mayer (2010) compared the language development of twin girls, only one of whom was born with profound deafness. The latter received two cochlear implants to aid her hearing and also received hearing therapy starting at age 1. You can see how this unique situation necessitated the use of a case study approach: It would be very difficult or even impossible to find large numbers of twins such as these two! The researchers spent two years thoroughly investigating this case using naturalistic observations in the home, formal language assessments, systematic observation in the lab, and structured interviews with their mother. What they found was that the implants and therapy were able to successfully reverse the early language delay experienced by the deaf twin. Remarkably, within about two years, the language development of the deaf child had caught up with her non-deaf sister.

☆ Student Spotlight: Case Studies ☆

For children with autism, a popular treatment is intensive behavioural intervention (IBI). In Ontario, a series of benchmarks were established in order to monitor the progress of treatment using IBI. Ksusha Blacklock, working under the supervision of Dr. Adrienne Perry, decided to examine whether the information contained in current clinical files for autism patients included enough information to evaluate these new benchmarks, and whether these benchmarks could be validated as a useful guideline for evaluating progress. In order to do so, six different case studies were conducted, based on a close examination of clinical files for child patients. Their findings were published in the *Journal on Developmental Disabilities* (Blacklock & Perry, 2010).

Sometimes, case studies of people with particular types of brain damage can allow also researchers to test hypotheses. For one example, consider research that a graduate student at York University, Donna Kwan, and her colleagues conducted (Kwan, Craver, Green, Myerson, & Rosenbaum, 2013). Previous research had shown that imagining future events activates a particular part of the brain: the hippocampus. This suggests that the hippocampus might be required in order to imagine the future. However, past case studies had resulted in mixed findings: Only some of those with hippocampal damage had difficulty imagining future experiences. Kwan and her colleagues used four case studies considered simultaneously to test the hypothesis that a fully functioning hippocampus is *not required* to imagine future outcomes. What they found was that the four people with hippocampal damage were able to consider the future to the same extent as those without any brain damage. This research provided evidence that hippocampal damage doesn't necessarily prevent all types of future-oriented thought. Moreover, because the same pattern was found in four different people (i.e., was replicated), the hypothesis received more support than if just one person was studied.

Case studies are valuable, and sometimes necessary, for learning about conditions that are rare or unusual. They can provide unique data about what is humanly possible for a range of psychological phenomena, such as memory, language, or decision-making. Insights gained through a case study may also lead to the development of hypotheses that can be tested using other methods. Nonetheless, extreme caution must be taken when interpreting the results of a case study, as it is inappropriate to generalize the results from one case to the population. In other words, we don't know if what holds true for one person also holds true for most people in the general population. This is an issue of *external validity* (i.e., the extent to which results can be generalized from the sample to the population from which it is drawn), which will be discussed in [Chapter 14](#).



Research

LO5 Archival

Archival research involves using previously compiled information to answer research questions. In this case, the researcher does not collect any original data. Rather, he or she analyzes existing data such as public records (e.g., number of divorce petitions filed), newspaper articles, blog posts, census data, public social media posts, and data published online for public access (e.g., NBA game statistics). An archival analysis can even involve studying published reports by other researchers (i.e., a *meta-analysis*; [Chapter 14](#)). The type of data extracted from archival sources differs depending on whether the researcher is using a qualitative approach (e.g., identifying common themes) or a quantitative approach (e.g., statistically analyzing numerical data). Three sources of archival research data are census data in the form of statistical records, survey archives, and written records.

Census Data or Statistical Records

Statistical records are collected by many public and private organizations. For example, Statistics Canada maintains the most extensive set of Canadian census records available to researchers, and it also makes these data available for analysis. Some of the many other sources of statistical records include provincial-level vital statistics (e.g., birth and marriage records), test score records by the Educational Testing Service (which administers standardized tests such as the Graduate Record Exam), the Centers for Disease Control and Prevention in the U.S., and even city-level data.

Many forms of public records can be used as sources of archival data. For example, Anderson and Anderson (1984) investigated the relationship between how hot the weather is and incidences of violent crime by drawing on records for the average monthly temperature and violent crime statistics for two U.S. cities. Data on both variables are readily available from agencies that keep these statistics. This is also true for Canada. As an example, Moulden and colleagues (2010) obtained access to a Royal Canadian Mounted Police database in order to describe the characteristics of teachers who have committed sexual offences against youth.

☆ Student Spotlight: Archival Research ☆

The Student Spotlight from [Chapter 1](#) about descriptive research presents a study on the prevalence of major depressive disorder in Canada. Because this study looked at data from the Canadian Community Health Survey, a national survey conducted by Statistics Canada that polls a representative sample of Canadians, it relies on pre-existing data collected for a separate purpose. Thus, it is an example of archival research. You can read more about their methodology in the journal *Canadian Psychology* (Knoll & MacLennan, 2017).

All major sports leagues keep and publish extensive records on many aspects of every game. These statistics are available to everyone and can be used to test some interesting hypotheses. For example, Tom Stafford (2018) examined data from over 5.5 million games of chess from international tournaments to see if female chess players underperform when matched against male opponents: This would provide evidence for

stereotype threat (i.e., when the activation of a stereotype impairs performance for the stereotyped person). In contrast to some past studies with smaller samples (e.g., Rothgerber & Wolsiefer, 2014), Stafford found just the opposite: Female players actually perform better than expected when playing against male opponents. As a Canadian example, Gee and Leith (2007) used archives from the National Hockey League (NHL) to test the hypothesis that players born in North America are more aggressive than players born in Europe. Despite earning, on average, the same number of points per season, North American-born players received more aggressive penalties than European-born players. As European-born players gained NHL experience, however, they tend to receive more aggressive penalties than European-born rookies. This suggests that aggressive acts are learned through experience playing hockey in North America. However, this is only one way to interpret the data. It is also possible that European players who do not like the aggressive style of North American hockey opt to leave the NHL, with only those who enjoy this style remaining in the league over time. Like any non-experimental finding, we must be careful to avoid inferring causality and consider all possible interpretations.

Survey Archives

Survey archives are a repository of data from surveys that is available to any researchers who wish to analyze them. For example, many major polling organizations make the data from their surveys available online. In addition, over 700 universities worldwide, including many Canadian schools, are part of the Inter-university Consortium for Political and Social Research (ICPSR), which also makes survey data freely available. Other very useful, publicly available, datasets include the World Values Survey (WVS; examined in [Chapters 12](#) and [13](#)) and the General Social Survey (GSS). The GSS is a series of surveys originally begun by the University of Chicago (now funded by the National Science Foundation), with a version created in Canada in 1985, funded by Statistics Canada. Each survey includes over 200 questions covering a wide range of topics, including attitudes, life satisfaction, health, religion, education, age, gender, and race. Using data from the Canadian GSS, Eagle (2011) investigated religious service attendance from 1986 to 2008. This study found a 20 percent overall decrease in weekly and monthly attendance, which was

accompanied by a 13 percent increase in the number of people reporting no religious affiliation or belief, coupled with a drop in the number of self-reported Catholics attending religious services (particularly in Quebec). One of the strengths of using the GSS is that it polls people from across Canada (aspiring toward a nationally representative sample). These sorts of archives are often extremely important, because most researchers do not have the financial resources to conduct surveys of randomly selected national samples, which are crucial for making descriptive claims about specific populations ([Chapter 7](#)). Archives allow researchers to access these difficult-to-collect samples to test their ideas.

Table 6.2 Publicly available data

There are many sources for publicly available data that anyone can download and analyze. Why not take a look at some of these, download some data, and see if you can answer some simple questions using a spreadsheet program?

City of Toronto Open Data Portal	https://portal0.cf.opendata.inter.sandbox-toronto.ca/
Academy Award Nominees and Winners from 1927– 2010	https://www.agpdata.com/awards/oscar
Government of Canada Open Data Portal	https://open.canada.ca/en/open-data
A Searchable Collection	https://www.kaggle.com/datasets

of Public
Datasets
Centers for
Disease
Control and
Prevention https://www.cdc.gov/nchs/data_access/ftp_data.htm



TRY IT OUT!

Using one of the publicly available datasets listed in [Table 6.2](#), try downloading and analyzing some public data to answer some simple questions. For example, visit the Government of Canada Open Data Portal and type “animal” into the search box. One of the first hits should be for *Animal Registrations and Transfers*. Click the link to access this document and then download the data in .CSV format (in which commas separate columns of data) by clicking “Access.” View the *Animal Registrations and Transfers Data Dictionary* to learn what all the variable names mean. Now load these data into a spreadsheet program and take a look at it (e.g., Google Sheets, Microsoft Excel, OpenOffice). Can you figure out how many registered pure-bred dogs are male and how many are female? (Hint: Try sorting by the “major commodity” column, then using a formula to calculate a sum for the relevant cells.) How many alpacas are registered in your province compared to a neighbouring province? Which province has registered the most goats?

Page 119

Written Records and Mass Media

Written records are stored documents, including diaries and letters that have been preserved by historical societies; public documents, such as speeches by politicians; and ethnographies, which are detailed descriptions of other cultures written by anthropologists (e.g., see the Human Relations Area Files: www.yale.edu/hraf). Examples of mass media records include books, magazine articles, movies, television programs, newspapers, and

Websites. Basically, any cultural product that is archived and accessible can be a potential source of data. Think about how much information is stored on the Internet: every status update and post on a social media platform, every photo uploaded, every e-mail sent, and every comment left on an article. In addition, behind the scenes, various forms of meta-data are being collected constantly by companies. This includes the terms entered into a search engine, location data based on mobile phone use, date and time stamps for every click, as well as time spent on each application, web page, and phone call. All of this information can potentially be used to answer research questions. In [Chapter 3](#), we discussed some ethical considerations related to large-scale datasets in our emerging “big data” society. Here we focus on their use as archival records of human behaviour, which can then be researched and analyzed. This wealth of data—and increasingly sophisticated ways of analyzing it—is fuelling the emergence of a sub-field in psychology known as *psychoinformatics* (Harlow & Oswald, 2016; Markowitz, Blaszkiewicz, Montag, Switala, & Schlaepfer, 2014; Yarkoni, 2012).

Relying on existing archives of written documents and mass media can be a rich source of data for non-experimental research ([Chapter 4](#)). These records can be used to study such complex behaviours as political action. For example, analyzing the cognitive structure embedded in politicians’ speeches has helped to understand and even predict decisions regarding war versus peace (see Suedfeld, 2010, for a review). Records can also be used to understand mental illness. Smartphone data such as physical movements (via GPS) and the usage of apps can be used to track the symptoms of major depressive disorder and Internet addiction (Markowitz et al., 2014). Simple text records can even tell us about emotions and how societies view gender, by studying the different emotions conveyed in newspaper birth announcements (Gonzalez & Koestner, 2005). Because mass media are products of culture, they reflect cultures and so can be studied in order to learn about a culture. One Canadian example involved investigating the content of tweets regarding CBC’s *Hockey Night in Canada*. This *content analysis* revealed several themes related to Canadian culture and the corporatization of hockey (Norman, 2012). Moreover, how we use language often reveals important elements of the self. This was demonstrated by Yarkoni (2010), who analyzed how words were used in a

sample of 694 blogs, finding that language use was linked to personality traits. In all these cases, previously existing records formed the data that were analyzed to learn something new about human behaviour.

Working with Archival Data: Content Analysis and Interpretation

[Content analysis](#) is the systematic analysis of existing archives, such as written documents and mass media records like movies and television shows (Weber, 1990). In order to analyze the content of these records, a coding scheme is used to quantify the information presented (like those described in the systematic observation section [earlier](#)). Sometimes the coding scheme is quite simple and straightforward. You could, for example, code how many people belong to different types of groups on Facebook. In other instances, the researcher must work to carefully define the coding categories, such as what would count as an instance of aggression in a movie (e.g., attempting physical harm, verbal insults). Whenever coding is conducted, raters must be trained on how to use the coding scheme appropriately, especially when categories might be nuanced or complex. Inter-rater reliability coefficients are always computed to check whether there is sufficiently high agreement among the raters.

There are also software solutions to aid with content analysis. For example, the computer program Linguistic Inquiry and Word Count (LIWC; Pennebaker, Booth, & Francis, 2007) counts the frequency of words that appear in texts, based on different categories. This program can tell researchers whether one group of texts (e.g., comments on a post) contains more words associated with negative emotions compared to another group of texts (e.g., the posts themselves). The words people use, and how they arrange them, can be examined to help reveal psychological processes (for a discussion see Tausczik & Pennebaker, 2010; for an example, see Yarkoni, 2010). Software can often help when analyzing many texts with lots of words, as manually reading and interpreting so much might not be feasible. However, these programs can be rather simple (e.g., simply identifying whether a word is present or not, and not considering its meaning within a context). Nuanced coding and interpretation of texts and

mass media will continue to require human thought (Lewis, Zamith, & Hermida, 2013). Software can also be used to analyze mass media archives. For example, James Cutting and Ayse Candan (2015) digitized over 10,000 films and examined the average length of a single shot in each (i.e., a continuous and uninterrupted image, also known as the length between cuts). In this way, they were able to discover that shots are getting shorter and shorter, which means that the pace of popular films is getting faster and faster with an increased number of cuts.

Archival data allows researchers to study interesting questions, some of which could not be studied in any other way. Based on these investigations, hypotheses can be tested and new hypotheses may also be generated. Archival data are a valuable supplement to more traditional data-collection methods, because they represent spontaneously occurring real-world behaviour. As a result, these studies offer an enhanced ability to generalize to the real world (i.e., possess good *external validity*; [Chapter 14](#)). That said, all research approaches have both strengths and weaknesses, and archival data are no different. There are at least three challenges associated with the use of archival data. First, the desired records may be difficult to obtain. Data can be placed in long-forgotten storage places, they may have been destroyed, or they may be owned by companies or institutions that aren't willing to share them (or perhaps not without charging an exorbitant fee!). For example, Bender and colleagues (2011) found that they could not analyze the demographic characteristics of different types of Facebook groups because of the privacy settings associated with these groups ([Chapter 3](#)). Second, for archival research we can't control what data were collected or how they were recorded. As a result, we can never be completely sure of the accuracy of the information that was collected by someone else. Third, as this work is non-experimental, alternative explanations for observed relationships exist, and so we cannot make causal claims about these associations.

Page 121

Table 6.3  TEST YOURSELF!

Read each scenario and determine whether the study used a case study, naturalistic observation, systematic observation, or archival research design.

Scenario	Design
Researchers conducted in-depth interviews with the few survivors of a tragic plane crash to better understand the psychological impact of these accidents.	
Researchers recorded how long it took drivers to back out of a parking space, also noting whether another car was waiting for that space or not.	
The contents of listings for a common product on eBay were coded, along with the time it took for the item to sell.	
The researchers conducted numerous interviews with a person born without a cerebellum (a large part of the brain), and administered numerous tests of physical and cognitive ability.	
Researchers examined gun ownership rates and the incidence of murder across 125 countries.	
Researchers studied recycling behaviour in city parks by making detailed field notes while posing as typical park-goers.	

This chapter explored several important observational techniques that can be used to study a variety of questions about behaviour, based on a quantitative or qualitative perspective. (Check out [Table 6.3](#) to gauge your understanding of this chapter.) In the next chapter, we examine a common way of finding out about human behaviour—simply asking people to tell us about themselves.



Illustrative Article: Observational Methods

Happiness, according to Aristotle, is the most desirable of all things. In the past few decades, many researchers have been studying predictors of happiness in an attempt to understand the construct.

Mehl, Vazire, Holleran, and Clark (2010) conducted a naturalistic observation on the topic of happiness using electronically activated recorders (a device that unobtrusively records snippets of sound at regular intervals, for a fixed amount of time). In this study, 79 undergraduate students wore the device for 4 days; 30-second recordings were made every 12.5 minutes. Each snippet was coded as having been taken while the participant was alone or with people. If the participant was with somebody, the recordings were also coded for “small talk” and “substantive talk.” Other measures administered were well-being and happiness.

First, acquire and read the article:

- Mehl, M. R., Vazire, S., Holleran, S. E., & Clark, C. S. (2010). Eavesdropping on happiness: Well-being is related to having less small talk and more substantive conversations. *Psychological Science*, 21, 539–541. doi:10.1177/0956797610362675

Page 122 Then, after reading the article, consider the following:

1. Is the basic approach in this study qualitative or quantitative?
2. Is this study an example of concealed or non-concealed observation? What are the ethical issues present in this study?
3. Do you think that participants would be reactive to this data collection method?
4. How reliable were the coders? How did the authors assess their reliability?
5. How did the researchers operationally define *small talk*, *substantive talk*, *well-being*, and *happiness*? What do you think about the quality of these operational definitions?

6. Does this study suffer from the problem involving the direction of causation? How so?
7. Does this study suffer from the third-variable problem? How so?
8. Do you think that this study included any confounding variables? Provide examples.
9. Given the topic of this study, what other ways can you think of to conduct this study using an observational method?

Study Terms

Test yourself! Define and generate an example of each of these key terms.

- archival research (p. 117).
- case study (p. 115).
- coding scheme (p. 113).
- concealed observation (p. 110).
- content analysis (p. 120).
- naturalistic observation (p. 108).
- participant observation (p. 110).
- qualitative approach (p. 107).
- quantitative approach (p. 107).
- systematic observation (p. 112).

Review Questions

Test yourself on this chapter's learning objectives. Can you answer each of these questions?

1. Contrast the major differences between qualitative and quantitative approaches to research. Describe how a researcher taking either perspective might use naturalistic observation.
2. What is naturalistic observation? How does a researcher collect data when conducting naturalistic observation research?
3. Distinguish between participant observation and non-participant observation. Distinguish between concealed and non-concealed observation. Compare the pros and cons for using each kind of observation.
4. What is systematic observation? What makes the data from systematic observation primarily quantitative?Page 123
5. What is a coding scheme? What are some important considerations when developing a coding scheme? How does it differ from a content analysis?
6. What is a case study? Why are case studies used?
7. What is archival research? What are the major sources of archival data?
8. Consider all the observational techniques described in this chapter. What limitations do they have in common? What unique limitations or challenges does each have?

Deepen Your Understanding

Develop your mastery of these concepts by considering these application questions. Compare your responses with those from other people in your study group.

1. Suppose you are interested in how a parent's alcoholism affects the life of an adolescent, and that you are open to considering both qualitative and quantitative approaches.
 1. Develop a research question best answered using quantitative techniques, then develop another research question better suited to qualitative techniques. For example, a quantitative question could be, "Are adolescents with alcoholic parents more likely to have criminal records than adolescents with non-alcoholic parents?" In contrast, a qualitative question could be, "What issues do alcoholic parents pose for their adolescent's peer relationships?"
 2. What techniques discussed in this chapter could you use to collect data that would help you answer those questions? What kind of data would you seek?
2. Devise a simple coding scheme to do a content analysis of a set of photographs. You could use your own uploads (e.g., Instagram or Facebook) or a publicly available collection (e.g., image search for "powerful"). Begin by examining the photographs to identify the content dimensions you wish to use. For example, if all the photographs in your set of images have people in them, you can use or adapt the list below as your coding scheme. Apply the coding to many photographs (e.g., 20) and describe your results. What hypotheses can you develop based on your results?

Criterion	Photo	Photo	Photo
	1	2	...

Criterion	Photo	Photo	Photo
	1	2	...
Number of people depicted (e.g., 1, 2)			
Facial expression (e.g., smiling, crying); <i>repeat per person</i>			
Activity (e.g., reading, hiking); <i>repeat per person</i>			
Indoors or outdoors			
Occasion (if known; e.g., birthday)			
Type of objects (e.g., drink, book)			
Geographic location (if known)			
Posed or candid			

Survey Research: Asking People about Themselves



©kajornyot/Getty Images

Like this common treeshrew, most people have a lot to say, so gathering information about others can be as simple as asking them.

Learning Objectives

Keep these learning objectives in mind as you read to help you identify the most critical information in this chapter.

By the end of this chapter, you should be able to:

1. [LO1](#) Discuss reasons for conducting survey research.
2. [LO2](#) Identify features to consider when writing questions for questionnaires and interviews, including defining research objectives, question wording, and response options.
3. [LO3](#) Describe different ways to construct questions, including open-ended questions and closed-ended questions.
4. [LO4](#) Compare two ways to administer surveys: written questionnaires and verbal interviews.
5. [LO5](#) Explain the relationship between sample size and precision.
6. [LO6](#) Describe the ways that samples are evaluated for potential bias, including sampling frame and response rate.
7. [LO7](#) Compare and contrast three kinds of probability and three kinds of non-probability sampling techniques.

Page 125 [Survey research](#) uses questionnaires and interviews to ask people to provide information about themselves. This includes information such as their attitudes and beliefs, demographics (age, cultural background, etc.), past behaviours, and intended future actions. Researchers using both qualitative and quantitative approaches employ surveys to collect data. Based on our discussion from [Chapter 6](#), you may already suspect that the kinds of questions asked and what researchers do with the data will differ depending on which approach they are currently using. In the first half of this chapter, we learn about methods for designing and conducting surveys. In the second half, we discuss how to interpret the results from surveys, taking into account various factors such as sampling techniques.



LO1 Why Conduct Surveys?

A multitude of surveys are being conducted all the time. The results of surveys frequently appear in the news. For example, you might read about Statistics Canada reporting the results of a survey of mental health symptoms in the Canadian Forces. Or a university student newspaper might report the results of its own survey of student satisfaction with campus food options. During election times, the results of polls of eligible voters dominate the news cycle. Surveys are clearly a common and important method of studying behaviour.

As behavioural scientists, we're frequently interested in other people, why they do what they do, and what they're thinking. Sometimes the simplest way to learn about others is to just ask them about themselves! Surveys provide us with a methodology for asking people to tell us about themselves. They have become extremely important as society demands data about issues, rather than just intuition and anecdotes. As an example, university departments need data from graduates to help determine what changes should be made to the curriculum. If the survey data are collected in a responsible fashion, these data may be much more useful sources of information than individual anecdotes. Surveys can inform the policy

decisions of lawmakers and public agencies. For example, the McCreary Centre Society in British Columbia (www.mcs.bc.ca) surveys young people across the province about health issues. These data are then used by the government and schools to create effective policies and programs. In basic research, many important variables, including attitudes, current emotional states, and self-reports of behaviours, are easily studied using questionnaires or interviews.

Surveys can be used in many ways. Often they are used to provide a “snapshot” of how a particular group of people think and behave at a given point in time. Surveys can also be used to gather data for studying relationships among variables, as well as ways that attitudes and behaviours change over time or differ among different groups of people. For example, researchers have examined the relationship between academic average and the number of hours that high school students work outside of school (Dumont, Leclerc, & McKinnon, 2009). The sample consisted of 187 Grade 9 students and another 144 Grade 11 students from a high school in Quebec. Consistent with previous research in the U.S. (Steinberg & Dornbusch, 1991), the more students worked outside of school, the lower their academic average. However, unlike earlier research, these researchers observed this pattern only among the Grade 9 students. For the Grade 11 students, average grades were unrelated to the amount they worked. In addition, students who decreased the amount they worked from Grade 9 to Grade 11 experienced an increase in their academic average over the same period. As with any non-experimental study, we do not know the causal direction of this relationship ([Chapter 4](#)). But we have learned something valuable: that these variables can have different relationships for different groups of people. This study also highlights a major benefit to survey methods: Simply asking students about their work hours provides an efficient way to gather important data. Consider the alternative: It would be far too difficult to visit each student’s place of work and measure their work hours directly.

Page 126

☆ Student Spotlight: Survey Research ☆

While an undergraduate student at the Université de Montréal, Marie-Ève Boucher collaborated with Stéphanie Arsenault and Drs. Serge Lecours and

Frederick Philippe on a study of depression. They were curious about how parents socialize their children to think about emotions (e.g., viewing sadness in a negative light), and how this might predispose individuals to depression as adults. In order to research their ideas, they asked 140 undergraduate students to complete questionnaire measures of the socialization of emotions, attitudes towards sadness, and symptoms of depression. Intercorrelations among these measures were then examined to investigate their hypotheses. The results of this study appeared in the *Revue Européenne de Psychologie Appliquée* (the *European Review of Applied Psychology*; Boucher, Lecours, Philippe, & Arseneault, 2013).

The National Survey of Student Engagement (NSSE; <http://nsse.iub.edu>) offers an example of a very large survey. Every year since 2000, thousands of students from hundreds of universities across Canada and the U.S. have rated their schools. Because the questions are the same each year, it is possible to track changes over time on such variables as level of academic challenge, student–faculty interaction, supportive campus environment, and enriching educational experiences. Universities whose students complete the survey are given valuable data (both quantitative and qualitative) that can help guide administrative decisions about how to improve undergraduate education. Closed-ended questions, with only certain numerical responses allowed, provide quantitative data that can be tracked over time, or used to compare scores for one institution with similar institutions. Open-ended questions, which allow for a variety of responses including text, can be content analyzed for additional insight into what undergraduates think would enrich their education ([Chapter 6](#)) (Chambers, 2006).

When the same people are tracked and surveyed at two or more points in time, the design is sometimes called a *panel study* or a *longitudinal design* ([Chapter 11](#)). Each period of sampling data is sometimes known as a “wave,” so a “two-wave” panel study involves surveying people at two points in time. Panel studies allow for research questions about the relationship between one variable (measured first, at “time one”) and another variable measured later (at “time two”). For example, survey data from the Canadian National Longitudinal Survey of Children and Youth (NLSCY) were used to examine children’s aggressive behaviours (Arim,

Dahinten, Marshall, & Shapka, 2011). The NLSCY surveys new parents—and the children themselves when they are old enough—every two years on a variety of variables related to family life, health, and development. Data from this survey showed that children who felt more nurtured and loved by their parents right before puberty (10 years old for girls, 12 years old for boys) engaged in less aggression two years later. This was true for both direct aggression (e.g., fist fights) and indirect aggression (e.g., spreading rumours).

Survey research is also important as a complement to experimental research findings. Researchers from the University of Waterloo and the University of New Brunswick collaborated to investigate the phenomenon of repeating positive affirmations to oneself (e.g., “I’m a lovable person,” “I will win!”), like those often recommended in self-help books or in pop psychology (Wood, Perunovic, & Lee, 2009). These researchers first used a survey of undergraduates to confirm that people actually do use these kinds of affirmations and believe that they are effective. Only 3 percent of participants reported that they had never used affirmations, and over half said they used them daily. After establishing that these affirmations are common, the researchers conducted experiments to understand their effects. It turns out that whether these affirmations actually made people feel good depended on their self-esteem. People with low self-esteem who were asked to repeat positive affirmations actually felt *worse* about themselves, compared to a control group. In contrast, people with high self-esteem felt a little bit better about themselves after repeating affirmations, compared to the controls. This set of studies illustrates an important point: Multiple methods are needed to understand any behaviour. Surveys were an essential starting point to establish that this phenomenon actually exists, and the follow-up experiments were key in establishing the causal consequences of engaging in the phenomenon.

Page 127

Response Bias in Survey Research

An assumption that underlies the use of questionnaires and interviews is that people are generally willing and able to provide truthful and accurate answers. However, the degree to which this assumption is untrue for any particular survey undermines the value of survey research. It’s important to

note that if a minority of people provide inaccurate information, the data as a whole may remain valuable and informative (i.e., there is more variance associated with true answers relative to measurement error). Only when there is an excess of inaccurate information that overshadows or masks the true variance in scores do we have serious problems. Researchers have addressed this issue by studying possible biases in the way people respond. A *response set* is a tendency to respond to all questions from a particular perspective, or in a particular way, rather than provide answers directly related to the questions themselves. Thus, response sets can reduce the usefulness of data obtained from self-reports.

One of the most common response sets is called *social desirability*, or “faking good.” The social desirability response set leads an individual to answer in the most socially acceptable way—the way the person thinks “most people” would respond, or the way that would present that person in the most positive light. Issues of social desirability are probably most likely to arise when the questions concern a sensitive topic, such as violent behaviour, alcohol or drug abuse, sexual practices, or socially unacceptable opinions (e.g., racist or prejudiced attitudes). This is quite a difficult problem to deal with, but one option is to ask sensitive or controversial questions in a way that hides their main intent. So, instead of asking “How often do you physically hit others?” you might ask “How often do you encounter people who are just asking to be hit?” There are also some scales that have attempted to measure the extent to which people try to present themselves in a favourable way. For example, one measure looks at the extent to which people claim familiarity with fake “facts” (e.g., that “plates of parallax” is a real term; Paulhus, Harms, Bruce, & Lysy, 2003). This might help researchers detect which participants are most likely to use a social desirability response set. Another way to increase accuracy in survey reports of sensitive topics is to ensure privacy while completing the survey, by stressing confidentiality and anonymity, and increasing motivation to respond accurately (for details, see Tourangeau & Yan, 2007). For example, you could have a classroom of people complete a survey without including their name or any identifiable information, then have them drop it into a box so they can see there’s no way to know who gave which responses. There are also some advanced statistical methods for creating tangible incentives to respond accurately (Prelec, 2004). In one study, an incentive

method was effectively used to encourage researchers to honestly report if they had ever engaged in ethically questionable research practices (e.g., conducting many different analyses, but only reporting those that supported their hypothesis; John, Loewenstein, & Prelec, 2012). Respondents were informed that an algorithm would be able to detect the amount of dishonesty across the entire sample, and that more honest responding would result in a larger donation to charity (which was all true). This method succeeded in increasing the amount of honest reporting for this sensitive topic.

Page 128

We turn now to the major considerations for survey research: constructing the questions, choosing how to present them, and deciding how to sample people to survey. Although it is often preferable to try to find surveys that have already been used and validated by other researchers ([Chapter 5](#)), sometimes it is necessary to construct a survey if no good option already exists. Just as importantly, understanding these major considerations will also help you to critique the surveys you encounter, to diagnose whether they are potentially problematic.



LO2 Constructing Good Questions

A great deal of thought must be given to crafting questions for questionnaires and interviews. This section describes some of the most important features to consider when constructing questions.

Defining the Research Objectives

When constructing questions for a survey, the first thing the researcher must do is determine the research objectives: What do they wish to know? The survey questions must be tied to the research questions that are being addressed. Surveys get out of hand when researchers begin to ask any question that comes to mind about a topic without considering exactly what useful information will be gained by doing so. Because scientists are curious by nature, it can be easy to ask all sorts of questions that might have interesting answers, but it is important to narrow our focus as much as possible. It is potentially unethical to ask questions if you have no real use for the information, as the more questions that are asked, the greater the burden placed on the participant (i.e., the greater the cost) with potentially little benefit. Thus, researchers must carefully decide on which questions to

ask and how to ask them. There are three general types of content measured by survey questions (Judd, Smith, & Kidder, 1991).

Attitudes and Beliefs

Questions about attitudes and beliefs focus on the ways people evaluate, feel, and think about issues. Should more money be spent on mental health services? Are you satisfied with the decision process for changes in government policy? How capable would you evaluate your instructor to be? Is there a way to improve student knowledge regarding financial investments and the perils of debt? These questions tap into important aspects of how people see the world and the value that different things hold for them.

Facts and Demographics

Factual questions ask people to indicate things they know to be true about themselves and their situation. In most studies, asking for some demographic information (e.g., age, gender identity) is necessary to adequately describe your sample so that researchers can consider the population from which it was drawn (and might generalize to). Depending on the topic of study, questions about things like ethnicity, income, marital status, employment status, handedness, and number of siblings might be included. Obviously, if you are interested in making comparisons across groups, such as rural versus urban upbringing, you must ask the relevant information about group membership.¹²⁹

Other factual information you might ask will depend on the topic of your research. Each year, *Consumer Reports* magazine asks readers to report any repairs that have been necessary for many products, such as cars and electronic devices. In a study about health and quality of life, factual questions about illnesses, hospitalizations, and other medical information would be asked.

Behaviours

Other survey questions can focus on behaviours, either things people have done in the past or intend to do in the future. How many times last week did you exercise for 20 minutes or longer? Have you ever been so depressed that you called in sick to work? How many courses do you plan to take next year? These are all questions about actions and behaviours that might be of interest to a researcher.

Question Wording

Survey questions must be carefully written to maximize the likelihood of eliciting informative responses and avoiding numerous potential problems (Graesser, Kennedy, Wiemer-Hastings, & Ottati, 1999). A great many problems stem from difficulties understanding the question, including (a) unfamiliar technical terms, (b) vague or imprecise terms, (c) grammatical errors (in sentence structure, for example), (d) awkward phrasing that taxes working memory, and (e) embedding the question with misleading information (see [Table 7.1](#)). Studies have demonstrated that question wording can influence how people respond (Koss, 1992), which is why we must be very careful when choosing our question wording. Here is a question that illustrates some of these problems: Did your mother, father, full-blooded sisters, full-blooded brothers, daughters, or sons ever have a heart attack or myocardial infarction?

Table 7.1 Question Wording: What Is the Problem?



TEST YOURSELF! Read each of the following questions and identify which problem(s) applies to each.

Negative Unnecessary Double- Loaded
Wording Complexity Barreled

**Negative Unnecessary Double- Loaded
Wording Complexity Barrelled**

1. To what extent do you agree that professors should not be disallowed from taking daily attendance?

2. I enjoy studying research methods and spending time with friends on my weekends.

3. Do you support the legislation that would unfairly raise taxes for hard-working professors?

4. I would describe myself as attractive and unintelligent.

Negative Unnecessary Double- Loaded Wording Complexity Barrelled

5. Do you believe the relationship between mobile phone interactions and the consumption of convenience-oriented foods is orthogonal?

6. Are you in favour of the tyrannical boss's whim to cut our valuable lunchtime by 30 precious minutes?

Let's review some of the problems that arise in this one question alone. First, it is written in a way that overloads your memory, all thanks to the length of the question, the need to keep track of all those relatives, and the need to consider two different diagnoses for each relative. Second, the term "myocardial infarction" may be unfamiliar to many people. So how do you write questions to avoid such problems? The following considerations are important to keep in mind.

Unnecessary Complexity

Survey questions should be as simple as possible, so that people can understand and respond to the questions easily. Avoid jargon and technical terms that most people wouldn't understand. You should aim to write your

questions in such a way that as many people as possible will understand them, even those who don't have an advanced degree or a lot of experience with the language. In some cases, it may be necessary to define a term or describe an issue prior to asking the question. For example, before asking whether someone approves of Bill S-4, you will need to provide a brief description of what the legislation proposes.

Double-Barrelled Questions

Avoid "double-barrelled" questions that ask two things at once, as this makes any one answer ambiguous and hard to interpret. An example of a double-barreled question is "To what degree do you enjoy the expense and excitement of sky-diving?" Based on this question, it's hard to know how to reply, especially if you enjoy the excitement but not the expense (or vice versa, although that's hard to imagine). As a result, we also don't know how to interpret a reply based on a rating scale. Are people indicating their agreement with regards to excitement, expense, or both? The solution to this issue is that if you are interested in more than one issue, ask it using more than one question (e.g., one question about excitement, and one about expense).Page 130

Loaded Questions

A loaded question is written in such a way as to try to bias people's response toward a particular answer. For example, the question "Do you favour eliminating the wasteful excesses in the public school budget?" includes the assertion that the budget includes wasteful expenditures that should obviously be eliminated. You can tell what sort of answer the question writer is hoping to elicit. It is also quite easy to see how a non-biased wording of this question would likely result in different answers, such as "Do you favour reducing the public school budget?" One way to detect a loaded question is to keep an eye out for emotionally charged words, words that are meant to create an emotional reaction or a particular evaluation. These types of questions might succeed in influencing the way that people respond, resulting in biased conclusions.

Negative Wording

It is also important to avoid framing questions in the negative, as this can confuse respondents as to how to correctly indicate their preference. For example, let's consider this negatively phrased question: "Do you believe that the city should not approve the proposed women's shelter?" If you state that you agree with this statement, it means you are disagreeing with the proposal, which can be confusing to understand. Empirically, studies have demonstrated that these negatively worded items reduce scale *reliability* and *validity*, as a result of the confusion they cause (e.g., Woods, 2006; [Chapter 5](#)). Removing this negative wording results in a far better, more easily comprehended question: "Do you believe that the city should approve the proposed women's shelter?"Page 131

“Yea-Saying” and “Nay-Saying”

When you ask several questions about a topic, a respondent may either agree (say “yea” or yes) or disagree (say “nay” or no) with all of the questions. These tendencies are called [“yea-saying” or “nay-saying” response sets](#). Agreeing with all the questions (“yea-saying”) is also known as having an acquiescence bias: a bias toward accepting the statement. Although it is possible that the respondent may in fact truly agree with each item, it's also possible that the person is simply indicating (dis)agreement without even considering the question itself, just to complete the survey quickly. One way to detect this response set is to word the questions so that true agreement with most items is unlikely. For example, a measure of loneliness (e.g., Russell, Peplau, & Cutrona, 1980) phrases some questions so that agreement means the respondent is lonely (e.g., “I feel isolated from others”) and others with the meaning reversed, so that disagreement indicates loneliness (e.g., “I feel part of a group of friends”). Someone who truly feels lonely is expected to agree with the former item but disagree with the latter. As a result, consistent (dis)agreement may indicate “yea-saying” or “nay-saying.” Although the logic is sound, recent research on questionnaire design has demonstrated that including a few reversed items tends to decrease scale reliability (Roszkowski & Soven, 2010).

Yea-saying and nay-saying are a form of [inattentive responding](#). This refers to people providing answers without even really thinking about them. A person who is not paying any attention to the questions might agree or

disagree with all the answers, or they might simply respond carelessly. If they do the latter, including some reversed items will not help you detect this issue. There are, however, much better solutions, such as including questions that request a particular response (e.g., “For this question, please answer with the second option from the left”; Marjanovic, Struthers, Cribbie, & Greenglass, 2014). This clever solution, the *Conscientious Responders Scale*, was actually devised by a Canadian researcher while still a graduate student in Toronto (Marjanovic et al., 2014). More sophisticated statistical methods for detecting inattentive responding also exist (Curran, 2016; Marjanovic & Holden, 2019).

☆ Student Spotlight: Constructing Surveys ☆

Have you ever suffered from insomnia? Having difficulties with sleeping, be it getting to sleep or staying asleep, is a very serious problem. In order to study it and better understand it, we need tools for measuring it and why it might occur. L. Odell Tan, working under the supervision of Dr. Thomas Hadjistavropoulos (University of Regina) and Dr. Ying MacNab (University of British Columbia), developed a self-report measure of catastrophic thinking regarding sleep-related issues. With the help of over 750 participants, they managed to develop an 18-item scale that includes subscales for rumination, magnification, and helplessness. Their scale is known as the Catastrophic Thoughts about Insomnia Scale (CTIS), and it is hoped that this survey could be used as a form of assessment to guide treatment for insomnia patients. You can read more about their scale, and the process they followed to develop and validate it, in the journal *Cognitive Therapy and Research* (Tan, Hadjistavropoulos, & MacNab, 2017).



LO3 Responses to Questions: What Kind of Data Are You Seeking?

Closed- versus Open-Ended Questions

There are two main types of questions: those that limit the responses possible, known as closed-ended, and those that don't limit the possible responses, known as open-ended. With *[closed-ended questions](#)*, a limited number of response alternatives are given, like with a multiple-choice question. With *[open-ended questions](#)*, respondents are free to answer in any way they like. The exact same question can be open-ended or closed-ended, depending on what responses are allowed. Thus, you could ask, "What is the most important thing children should learn to prepare them for life?" followed by a list of possible answers (making this a closed-ended question), or you could simply provide an empty box for people to write in any answer they choose (in the form of an open-ended question).

Closed-ended questions are easier to code, by converting the response options to numerical data that can easily be analyzed. In contrast, open-ended questions require time to code the responses using content analysis or other qualitative methods ([Chapter 6](#)). Coding these responses is an effortful process. Sometimes,

a respondent's response cannot be categorized at all because the response doesn't make sense. In this way, open-ended questions require more time and resources to analyze. That said, open-ended questions are a richer source of information than closed-ended questions, because of the wide variety of responses possible. Moreover, by not limiting the respondent to a set of possible answers, open-ended questions avoid missing out on possibly useful information. These types of questions are very useful for researchers who need to know what people are thinking and how they view their world. Open-ended questions are also very useful when first researching a new topic, when you don't know enough to make educated assumptions about what sorts of answers people might provide. Closed-ended questions, in contrast, are more likely to be used when the variables under study are well-known and well-defined. As you may have suspected, closed-ended questions tend to be favoured by those using a quantitative approach, whereas open-ended questions are used by both, yet favoured by those using a qualitative approach.

Using either closed- or open-ended questions can sometimes lead to different conclusions. Let's consider our original question about preparing children for life (Schwarz, 1999). When this was posed as a closed-ended question, 62 percent of people chose the response "To think for themselves," from among a set of possible responses. However, when this was asked as an open-ended question in which people could reply in any way they wished, only 5 percent gave this answer. This finding highlights the need to have a good understanding of people's natural responses to a topic when asking closed-ended questions. When entering a new research area, a qualitative approach can be a useful starting point for developing the appropriate closed-ended questions for a new survey measure.

Rating Scales for Closed-Ended Questions

With closed-ended questions, there are a set number of response options. A [rating scale](#) asks people to provide judgments of "how much" for a dimension—for example, amount of agreement, liking, or confidence. Rating scales can have many different formats and use various measurement scales (e.g., ordinal; [Chapter 5](#)). The format that is used depends on factors such as the topic being investigated. Perhaps the best way to gain an understanding of the variety of formats is to look at a few examples. The simplest and most direct scale presents people with five or seven response alternatives, with labels defining either just the endpoints or all points of the scale. Consider the following examples:

Page 133

All dogs should be required to pass an obedience course before they can be called a “good boy” or “good girl.”

1	2	3	4	5	6	7
Strongly Disagree	Disagree	Slightly Disagree	Neither Agree nor Disagree	Slightly Agree	Agree	Strongly Agree

How confident are you that most cats love their owners?

0	1	2	3	4
Not at All Confident				Very Confident

Labelling Response Alternatives

The second example above provides labels for only the endpoints on the rating scale. Respondents decide for themselves the meaning of the other response alternatives. Although people are usually able to use such scales without difficulty, research shows that fully labelled scales are more reliable than are partially labelled scales (as in the first example above) (Krosnick, 1999). Labels help to more clearly define the meaning of each alternative and reduce the measurement error that arises from relying on each person’s individual interpretations of the response options ([Chapter 5](#)). The first example above represents a fairly standard way of labelling response alternatives for an agreement measure (Likert, 1932).

Number of Response Alternatives

In public opinion surveys, a simple “yes or no” or “agree or disagree” dichotomy can be sufficient. In more basic research, it is often preferable to have more alternatives to allow people greater opportunity to express themselves.

Researchers often choose 5- or 7-point scales, allowing a middle “neutral” option, as in the first example above. This type of scale assumes that the middle alternative is a “neutral” point, halfway between the endpoints. However, respondents sometimes use this middle point when they don’t know how to respond, which can cause problems for reliability and validity. Therefore, it is often recommended to offer an “I don’t know” or “Not applicable” option to participants. Another issue is whether to have an odd or even number of alternatives. It is possible to force participants to choose one side or the other by dropping the “Neither agree nor disagree” option and keeping a 6-point rather

than a 7-point scale, although this decision can substantially impact the way participants respond to a questionnaire (Nowlis, Kahn, & Dhar, 2002).

Sometimes a perfectly balanced scale may not be possible or desirable. Consider a scale that is commonly used on forms asking a professor to rate an undergraduate student who is applying for a job or graduate program. This particular scale asks for comparative ratings of students:

In comparison with other graduates, how would you rate this student's potential for success?

Lower	Upper	Upper	Upper	Upper
50%	50%	25%	10%	5%

Notice that most of the alternatives are asking people to make a rating in terms of the top 25 percent of students. This is done because students who apply for such programs tend to be very bright and motivated, and so professors rate them favourably. The wording of the alternatives attempts to force the respondents to make finer distinctions among generally very good students. Page 134

Labelling alternatives can be particularly tricky when asking about the frequency of some behaviour. For example, for the question, "How often do you exercise for at least 20 minutes?", what kind of scale should you use to collect responses? One possibility is to list: (1) never, (2) rarely, (3) sometimes, and (4) frequently. These terms convey your meaning, but they are rather vague. Here is a more precise set of alternatives (Schwarz, 1999): (1) less than twice a week, (2) about twice a week, (3) about four times a week, (4) about six times a week, (5) at least once a day. This scale is known as a *high-frequency scale*, because most of the options measure a behaviour done very frequently. But what if the behaviour in question does not happen as frequently? Let's consider the question, "How often do you pay more than \$50 for a meal?" In this case, an appropriate response scale might look more like this: (1) less than once per year, (2) less than four times a year, (3) about every other month, (4) about once a month, (5) about once every two weeks. This is known as a *low-frequency scale*, because the behaviour in question doesn't happen very frequently. You can imagine how using a high-frequency scale for this question about expensive meals might not be appropriate, especially for your typical undergraduate!

Labels should be chosen carefully because people may interpret the meaning of the scale differently, depending on the labels used. If you were actually asking the

cost-of-meal question, you might decide on alternatives different from the ones described here. Moreover, your choice should be influenced by factors such as the population you are studying. If you are studying people who are very wealthy, you will be more likely to use a high-frequency scale than you would if you were studying people who are not.

Graphic Rating Scale

Another response scale option is a [graphic rating scale](#), which asks respondents to make a mark along a continuous 100 mm line, anchored with descriptions at each end.

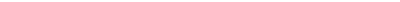
- How enjoyable is your Research Methods course?
 - Not very enjoyable _____ Very
enjoyable

The researcher then measures where the mark was made, converting this to a score that ranges from 0 to 100. Current survey software can also calculate the score for graphic rating scales automatically.

Semantic Differential Scale

The *semantic differential scale* is a way to measure the meaning that people ascribe to concepts (Osgood, Suci, & Tannenbaum, 1957). Respondents rate concepts based on a continuum spanning two adjectives (one at each end), using multi-point scales (e.g., 7 points between the two adjectives). These concepts can be anything: people, objects, behaviours, ideas, etc. In the example below, a respondent who felt that wearing shorts outside during a Canadian winter was totally bad and not at all good would place a check mark on the line closest to “Bad.” If she felt that wearing shorts in the winter was a sign of strength, she would place a check mark on a line closer to “Strong.” Using multiple adjective pairs can help measure different aspects of attitudes toward a concept. So a person might believe that wearing shorts outside when it’s extremely cold is “Bad” and “Stupid,” but also associated with strength.

Wearing Shorts outside in the Canadian Winter

Good  Bad

Strong _____ Weak

Smart _____ Stupid

Page 135 Research on semantic differential scales shows that virtually anything can be measured using this technique. Ratings of specific things (emus), places (Halifax), people (emu keepers), ideas (guaranteed minimum income), and behaviours (break-dancing) can be obtained. Concepts rated using semantic differential scales are rated along three basic dimensions: The first and most important is *evaluation* (e.g., how something is valued, using adjectives such as good–bad, wise–foolish, kind–cruel). The second is *activity* (active–passive, slow–fast, excitable–calm). And the third is *potency* (weak–strong, hard–soft, large–small).

Non-verbal Scale

In some circumstances, a researcher might want to offer response options in the form of images instead of words or numbers. For example, young children may not understand the types of scales we've just described, but using images they are also able to give ratings. For example, you could ask children to "Point to the face that shows how you feel about bedtime."



Finalizing the Questionnaire

Formatting the Questionnaire

The final questionnaire, whether distributed on paper or online, should appear attractive and professional. It should be presented in a clear font and free of spelling errors. Respondents should find it easy to identify questions and their corresponding response options. Leave enough space between questions so that people don't become confused when reading the questionnaire.

Carefully consider the order of your questions. In general, it is best to ask the most interesting and important questions first to capture the attention of your respondents and motivate them to complete the survey. For one employee attitude survey, the highest return rates were observed when important questions were presented first and demographic questions were asked last (Roberson & Sundstrom, 1990). This strategy also handles participant fatigue quite well, by asking the important questions at the beginning when respondents are least tired, and the easiest questions at the end when they might be a bit fatigued (i.e., factual questions pertaining to demographics that someone can answer easily). Participants can easily tire of long surveys, so try to make your survey as concise as possible. When you have multiple questionnaires within a single survey, randomizing the order of these questionnaires can also ensure that no particular questionnaire is subject to more participant fatigue than any other. If one of your measures always appears near the end, and respondents are feeling pretty tired and unmotivated by this point, their lazy responding could increase the measurement error for this measure in particular. Along similar lines, randomizing the order of items within a questionnaire also helps to evenly distribute any effects of fatigue or order. Randomization of this sort is relatively easy to accomplish with online survey software. Page 136

Questionnaire formatting can also be optimized to improve response rates in online surveys (Vicente & Reis, 2010). The current evidence suggests

that it is best to present only a few questions per page, rather than having one long page that people must scroll down to see all the items. A progress indicator bar at the top of the screen can also help ensure completion, as long as participants perceive they are moving through the survey at a fast-enough rate. Using “radio buttons,” in which people select response options that are all visible, is preferable to drop-down menus for closed-ended items. Lastly, because shorter questionnaires produce higher response rates than longer ones, and to reduce participant fatigue, carefully consider the necessity of every question you plan on including. Several online options are available for presenting surveys, including Qualtrics ([qualtrics.com](https://www.qualtrics.com)), SurveyMonkey ([surveymonkey.com](https://www.surveymonkey.com)), or its Canadian subsidiary FluidSurveys ([fluidsurveys.com](https://www.fluidsurveys.com)), which has domestic servers (important for legal aspects of data privacy).

Refining Questions

Before actually administering the survey, it is a good idea to give the questions to a few people and have them “think aloud” while answering them. These participants might be chosen from the population being studied, or they could be friends or colleagues who can give reasonable responses to the questions. Ask them to tell you how they interpret each question and how they respond to the response alternatives. This procedure can provide valuable information you can then use to improve the questions.



Think about It!

By asking some people to “think aloud” while going through your survey, what sort of approach are you using? Qualitative or quantitative? Why is this approach useful for this purpose?



LO4 Administering Surveys

There are two main ways to administer surveys. In the first, the *questionnaire format*, respondents read the questions and indicate their responses on a paper or online form. In the second, the *interview format*, an interviewer asks the questions and records responses from the verbal interaction. In this section, we will explore unique issues raised by each method.

For either format, there are also a number of challenges that emerge when you measure the same participants repeatedly. In [Chapter 3](#), we discussed options for maintaining confidential records with such designs.

Additionally, changes in an institution or even society at large might influence responses recorded over a period of time. For example, a university might make a major policy change or experience drastic funding cuts one year, which would affect how we interpret scores that are recorded before versus after this change. Another major concern with measurement that occurs over time is that not all participants will stay in the study as time goes on: Some people will inevitably drop out of the study over time. In this way, a sample that starts out *random* (see [below](#)) or representative of a population can actually become biased over time, if it is not entirely

random who drops out versus who decides to stay. These issues are common to all studies that use a within-subjects design or have a *longitudinal* component. We will revisit these discussions in [Chapters 8](#) and [11](#).

Questionnaires

Questionnaires present respondents with questions in written format and answers are written out or typed into an online form. Questionnaires can be administered in person to groups or individuals, through the mail, on the Internet, or through mobile devices (e.g., smartphones, smart watches). A major benefit to using questionnaires is the low cost. They also allow the respondent to be completely anonymous, as long as no identifying information is asked (e.g., name, e-mail address) and Internet protocol (IP) addresses are not recorded in online surveys. This anonymity might encourage respondents to give honest answers to your questions. However, questionnaires require respondents to be both motivated and attentive enough to complete them honestly. Some populations may have difficulty reading and understanding questions. In these cases, it can be useful to administer the survey in person so that a researcher is there to clarify the questions if needed.

Personal Administration to Groups or Individuals

Often researchers are able to distribute questionnaires to groups of individuals. This might be a university class, parents attending a school meeting, people attending a new employee orientation, or people waiting in line. An advantage of this approach is that you have a captive audience that is likely to complete the questionnaire once they start it. Also, the researcher is present, so people can ask questions if necessary.

Mail Surveys

Paper surveys can also be mailed to individuals at a home or business address. One major drawback to this approach is the typically low rate of responding. These questionnaires can easily be placed aside and forgotten

among all the other tasks that people must attend to at home and at work. Even if people start to fill out the questionnaire, something may happen to distract them, they may have difficulty with a question, or they may become bored and simply throw the form in the recycling bin. Some methods for increasing response rates are described [later](#) in this chapter.

Internet Surveys

Administering questionnaires online is both easy and inexpensive, and this is one of the most common approaches for delivering surveys. Responses can be immediately downloaded for analysis, much easier than manually entering the data into a spreadsheet as is the case for many other forms of administering questionnaires. Researchers and major polling organizations sometimes build databases of people interested in participating in surveys who they can e-mail with an invitation to participate and, if appropriate, to forward the link to other potential participants. Online surveys can also be advertised using social networking sites and online forums. For example, if you are interested in a particular group of people who share an interest, you can advertise your survey in an online group dedicated to this interest. One advantage to using the Internet is that it is possible to target advertisements and posts to obtain very large samples of people with particular characteristics, such as marital status or occupational group. For example, one online study was able to reach a sample of over 650 people, all of whom have a family member or friend who suffers from hoarding behaviour, to find out how that illness burdens them (Tolin, Frost, Steketee, & Fitch, 2008).

Internet surveys can also be combined with other technologies to gather data. Hanson and Chen (2010) asked a sample of University of British Columbia students to complete an online survey about their day's stressors at the end of each day for a week. Participants then wore a special device to bed, which measured their sleep quality each night. Researchers learned that experiencing many hassles during the day predicted interrupted sleep quality among university students who had difficult childhoods, but this was not true for those who had experienced easier childhoods. Although asking participants to complete a daily paper-and-pencil measure was an option, using a website enabled researchers to confirm exactly when their

participants completed the survey. Technology played a vital role in this research. Page 138

One concern about Internet data collection is whether the results will be similar to what might be found using more traditional methods. However, researchers have found that data collected on the Internet are comparable to that collected in person (e.g., Howell, Rodzon, Kurai, & Sanchez, 2010; Gosling, Vazire, Srivastava, & John, 2004). One problem that remains with collecting data over the Internet is that the true characteristics of the respondents are ambiguous. To meet ethical guidelines, researchers will usually state that only people 16 years of age or older are eligible to participate, but this cannot be strictly controlled. People may also misrepresent their sex or ethnicity. We simply do not know if this is a major problem. For most research topics, it seems unlikely that people will go to the trouble of misrepresenting themselves on the Internet to a greater extent than they would with any other method of collecting data. For further consideration of ethical issues with Internet research, see Kraut and colleagues (2004).

Other Technologies

Researchers are continually exploring new technologies to assist with the collection of survey data. For example, consider studies seeking to sample people's behaviours or emotions over an extended period of time. Instead of asking people to remember how they felt or acted over the past week, researchers using computerized experience sampling contact participants at various times throughout the day on their cellphone or other mobile device and ask them to provide an immediate report of their current activities and experiences (e.g., Feldman-Barrett & Barrett, 2001). Sampling people's experiences throughout the day may result in a more accurate picture than relying on retrospective recall. For example, researchers have found that people report feeling more risk (e.g., of physical harm) in their daily lives when they report it via text message throughout their day than when they rely on their memory to complete a paper-and-pencil measure (Hogarth, Portell, & Cuxart, 2007).

Interviews

The fact that an interview involves an interaction between people has important implications. First, people are often more likely to agree to answer questions for a real person than to answer a mailed questionnaire. Thus, response rates tend to be higher when interviews are used, compared to questionnaires. The interviewer and respondent can establish a rapport that helps motivate the person to answer all the questions, rather than leave questions unanswered. An important advantage of an interview is that the interviewer can address any problems the person might have in understanding questions. Further, an interviewer can ask follow-up questions if needed to help clarify answers.

One potential problem in interviews is called *interviewer bias*. This term summarizes all of the different biases that can arise from the fact that the interviewer is a person interacting with another person. Thus, one potential problem is that the interviewer could subtly influence the respondent's answers by inadvertently showing approval or disapproval of certain answers. Personal characteristics of the interviewers (e.g., physical attractiveness, age, race) might also influence a respondent's answers. Interviewers may also have expectations that could lead them to "see what they are looking for" in the respondent's answers. Such expectations could bias their interpretations of responses or lead them to probe further for an answer from some respondents, but not from others (e.g., those with different cultural backgrounds; see also *experimenter expectancy effects* in [Chapter 9](#)). Careful screening and training of interviewers help to limit such biases.
Page 139

There are three methods of conducting interviews: face-to-face, by telephone, and with focus groups. Focus groups are mainly used for qualitative research, whereas face-to-face and telephone interviews are used for both qualitative and quantitative research.

Face-to-Face Interviews

Face-to-face interviews require that the interviewer and respondent meet to conduct the interview. The interviewer may travel to the respondent's home or office, or the respondent may go to the interviewer's office. Doing interviews face-to-face tends to be quite expensive and time-consuming.

Therefore, they are most likely to be used when the sample size is fairly small and when there are clear benefits to a face-to-face interaction.

Telephone Interviews

Interviews for large-scale surveys are commonly completed using telephone or equivalent online programs (e.g., Skype, FaceTime). Telephone interviews are less expensive than face-to-face interviews, and they allow data to be collected relatively quickly because many interviewers can work on the same survey at once. With a computer-assisted telephone interview system, the interviewer's questions are prompted on the computer screen, and the interviewer enters the data directly into the computer database, ready for analysis.

Focus Group Interviews

A *focus group* is an interview with a group of about six to ten people brought together for a period of usually two to three hours. Often the group members are selected because they have particular knowledge or interest in the topic. Because the focus group requires people to spend time and money travelling to the focus group location, there is usually some sort of incentive to participate (e.g., money or a gift).

In focus groups, the questions tend to be open-ended and are asked of the whole group. Group interaction is considered an advantage in this method: People can respond to one another, and one comment can trigger a variety of responses. The interviewer must be skilled in working with groups to facilitate communication and to deal with problems that may arise, such as one or two people trying to dominate the discussion or hostility between group members. The group discussion is usually recorded and may be transcribed. The recordings and transcripts are then analyzed to find themes and areas of group consensus and disagreement (i.e., a more qualitative approach; [Chapter 6](#)). Sometimes the transcripts are content analyzed to search for the frequency of certain words and phrases appearing (i.e., a more quantitative approach). Researchers usually prefer to conduct at least two or three focus groups on a given topic to make sure that the information gathered is not unique to one group of people. However,

because each focus group is time-consuming and costly, and provides a great deal of information, researchers typically don't conduct many groups on any one topic.

Interpreting Survey Results: Consider the Sample

Of the methods discussed in this book, surveys are used most frequently to gather responses from large groups of people and to make claims about the characteristics of these groups. These two features highlight some issues regarding interpretation centred around sample size and sampling techniques. Although some of those issues will be explored more generally elsewhere (e.g., [Chapter 13](#)), let's briefly examine them in the special context of surveys.

Population and Samples

A *population* is a set of people of interest to the researcher. This can be any group of people. For a large national election poll, for example, the population of interest might be all eligible voters in Canada. Or perhaps researchers are interested in the experience of Canadian university students, in which case the population of interest would be every university student in Canada. With enough time and money, a survey researcher could conceivably contact everyone in the population of interest.

With a relatively small population of interest (e.g., everyone in your immediate family), you might find it easy to study everyone in the population. In most cases, however, studying the entire population of interest would be a massive undertaking and is typically impossible or unfeasible. Because of financial and time constraints, researchers typically collect data on a sample of the population (i.e., a subset) in an attempt to learn something about the larger population from which that sample was drawn. With proper *sampling*, we can use information obtained from the respondents who were sampled to estimate characteristics of the population. Note that the more of the population that we manage to sample, the more confident we will be that the results of our sample represent the results we would observe were we able to test everyone in the population.

Imagine that you wish to poll your chess club, which has ten members, about a potential name change. You would likely feel more confident that the results of your poll represents the group's shared opinion if you hear from nine people as opposed to only three. In addition, statistical theory allows us to infer what the population is like, based on data obtained from a sample we will explore this idea further in ([Chapter 13](#)).

Confidence Intervals

When researchers make inferences, or estimates, about a population using a sample, they do so with a certain degree of confidence: Since we're only making informed guesses, we're never 100 percent certain. Whenever we see an estimate, we should wonder about the uncertainty that surrounds this estimate. One way to quantify this uncertainty is through a confidence interval. For example, an estimate for an average approval rating of 61 percent might also report a 95 percent confidence interval, between 58 and 64 percent. The true definition of a [*confidence interval*](#) is actually quite confusing and very often misunderstood, even by researchers who use them frequently (for the true definition, see Hoekstra, Morey, Rouder, & Wagenmakers, 2014). But, in practice, if a study is repeated, there is an 83 percent chance that the estimate observed in this replication will fall within the 95 percent confidence interval for the original study (Cumming, 2008). In this way, you can think of confidence intervals as a range of plausible values for direct replications. Confidence intervals tell us about the uncertainty around our estimates, with wider confidence intervals indicating greater uncertainty (i.e., a wider range of plausible values for what might be observed during a replication of the study).

To recap, the value we observe in a sample is our estimate of the value that exists in the population, with the confidence interval helping us to account for the error in this estimate (because it is based on a sample, or subset, of our population and not the whole population). The formal term for this error is [*sampling error*](#), although you may have seen the term *margin of error* instead. Recall that the concept of measurement error was discussed in [Chapter 5](#): When you measure a single individual on a variable, the obtained score deviates from the true score as a function of the measurement error. Similarly, when you study a sample, the obtained

estimate deviates from the true population value as a result of sampling error. The confidence interval gives you information about how much error likely exists. The topic of confidence intervals is revisited more broadly in [Chapter 13](#). For now, remember that narrower confidence intervals indicate more precise estimates with less sampling error and less uncertainty regarding these estimates.

Page 141



LO5 For More Precise Estimates, Use a Larger Sample

Larger sample sizes reduce measurement error, and therefore reduce the size of the confidence interval. Although the size of the interval is determined by several factors, sample size is key. In general, larger samples are more likely to yield data that accurately reflect the true population value, just like with our example regarding a chess club (above).

How large should the sample be? This is a complicated question with many possible answers, but one way to determine an ideal sample size is to calculate the size of the confidence interval based on the size of the population you are studying (Cumming & Finch, 2005). [Table 7.2](#) provides some examples of how sample size affects precision based on the population size. Notice that greater accuracy requires larger sample sizes, and that larger samples are also required the larger the size of the population. You can also see that sample size is *not* a constant percentage

of the population size. Many people incorrectly believe that proper sampling requires a certain percentage of the population. However, you can see in the table that the required sample size does not change much even as the population size increases from 5,000 to 100,000 or more. As Fowler (1984) notes, “A sample of 150 people will describe a population of 1,500 or 15 million with virtually the same degree of accuracy. . .” (p. 41), as long as the technique used to sample participants is appropriate (which we will discuss [below](#)).

Table 7.2 Sample Size Needed for Population Estimates at Three Levels of Precision (95% Confidence Level)

Size of Population	Precision of Estimate		
	±3%	±5%	±10%
2,000	696	322	92
5,000	879	357	94
10,000	964	370	95
50,000	1,045	381	96
100,000	1,056	383	96
Over 100,000	1,067	384	96



LO6 To Describe a Specific Population, Sample Thoroughly

Most researchers are interested in an entire population of interest, and so it is crucial that findings based on a sample can be generalized to this population. In other words, the research must have high *external validity*. Achieving external validity means first ensuring that the sample is highly representative of the population from which it is drawn. To create an unbiased sample, first you would randomly sample from a population that contains *all* people in the population of interest, with every person in the population having an equal chance of appearing in your sample. Second, you would contact and obtain completed responses from *all* people selected to be in the sample. In this way, although you have taken a subset, the data you end up with comes from a totally random selection of individuals from the population. Such standards are rarely achieved. Various non-random sampling methods (described later in this chapter) introduce different biases into the sample, but are used out of necessity when random sampling of populations is impossible.

Page 142
Regardless of the sampling method used, bias is introduced to the sample process based on the sampling frame and response rates. Let us consider these major sources of bias before considering a variety of sampling techniques in more detail.

Sampling Frame

The *sampling frame* is the actual population of people (or clusters of people) from which a random sample will be drawn, which is often a subset of the population of interest. Rarely will the sampling frame perfectly coincide with the population of interest—some biases will be introduced. The severity of these biases will impact the external validity of your results. If you define your population of interest as “residents of Saskatoon,” the sampling frame may be based on a list of home telephone numbers that you will use to contact residents between 5 p.m. and 9 p.m. This sampling frame excludes people whose schedule prevents them from being at home when you are making calls. Also, if you are using the telephone directory to obtain numbers, you will exclude people who have unlisted numbers, those who exclusively use a cellphone and do not have a home phone number, and those who do not have a telephone at all. When evaluating research, try to consider how well the sampling frame matches the population of interest. Often, the biases introduced are quite minor, although in some cases they can be consequential.

Consider the following example of how sampling frame can introduce highly consequential bias. For the 2013 British Columbia provincial election, all major election polling companies were shocked when they wrongly predicted the results; none of them had predicted that NDP leader Adrian Dix would lose to Liberal leader Christy Clark. Months of post-election analysis revealed that *sampling frame* was one of the key flaws. The sampling frame was defined as “eligible voters in British Columbia,” rather than “people who will vote in this election” (Reid, 2013). Voters aged 18 to 35 tend to support the NDP, but do not tend to vote. Because this age group represents a larger proportion of eligible voters compared to actual voters, they were overrepresented in the sample used to make predictions. Consequently, those predictions became biased and inaccurate.

Response Rate

The *response rate* in a survey is the percentage of people in the sample who actually complete the survey. If you send 1,000 questionnaires to a random sample of adults in your community and 500 are completed and returned to

you, the response rate is 50 percent. Response rate is one indicator of how much bias there might be in the final sample of respondents. Bias could be introduced if non-respondents differ from respondents in any number of ways, including age, income, marital status, and education. The lower the response rate, the greater the likelihood that such biases may distort the findings and in turn limit the ability to generalize the findings to the population of interest. In other words, the study will have low *external validity*, because the sample is composed of more (or less) of one type of person than is present in the entire population.

An example of response rate bias occurred in 2010, when the Canadian government cancelled the mandatory long-form version of the national census, making the extended questions optional. Data from 2011 showed lower response rates for these optional questions than with the former long-form census. Importantly, the decreases in response rate were uneven across different groups (e.g., across regions; Statistics Canada, 2013). Certain groups may be unrepresented entirely, and others may be heavily overrepresented. Some have argued that the sample no longer reflects the demographics of Canadians, thereby reducing the ability of agencies and businesses to provide appropriate services in particular regions (Cain & Mehler Paperny, 2013; Cain, 2013).Page 143

In general, mailed surveys have lower response rates than Internet and telephone surveys. With all methods, however, steps can be taken to maximize response rates. Researchers might motivate some people to respond by highlighting the importance of the survey and that their participation will make a valuable contribution. With mailed surveys, an explanatory letter can be sent a week or so prior to mailing the survey. Follow-up reminders, second mailings of the questionnaire, and including a stamped return envelope are often effective in increasing response rates. Even the look of the cover page of the questionnaire can be important (Dillman, 2000). With Internet surveys, follow-up e-mails are critical. With telephone surveys, people who aren't home can be called again, and people who can't be interviewed today can be scheduled for a call at a more convenient time. Sometimes an incentive may be necessary to increase response rates. Such incentives can include cash, a gift, a gift certificate, or a chance to win a prize in exchange for participation.



Techniques

There are two basic techniques for sampling individuals from a population: probability sampling and non-probability sampling. The sampling technique you use has implications for what conclusions you can draw. In *probability sampling*, each member of the population has a known and specific probability of being chosen. Probability sampling allows for representative samples, allowing the results from samples to be generalized to the population from which they were drawn. In *non-probability sampling*, we don't know the probability of any particular member of the population being chosen. This has implications for the generalizability of any results based on the sample. However, because of the difficulties of probability sampling, non-probability sampling is quite common and can be necessary in certain circumstances. ([Table 7.3](#) has a review of the advantages and disadvantages of each sampling technique.)

Table 7.3 Advantages and Disadvantages of Sampling Techniques

Sampling Technique	Example	Advantages	Disadvantages
<i>Probability sampling</i>			
Simple random sampling	A computer program randomly chooses 100 students from a list of all 10,000 students at a university.	Representative of population.	May cost more. May be difficult to get full list of all members of any population of interest.
Stratified random sampling	The names of all 10,000 university students are sorted by major, and a computer program randomly chooses 50 students from each major.	Representative of population.	May cost more. May be difficult to get full list of all members of any population of interest.
Cluster sampling	Psychology majors are identified at 100 schools all over Canada. Out of these 100 clusters, 10 clusters are chosen randomly, and every psychology major in each cluster is sampled.	Researcher does not have to sample from multiple lists of people in order to get a random sample.	May cost more. May be difficult to get full list of all members of any randomly chosen cluster.

Non-probability sampling

Sampling Technique	Example	Advantages	Disadvantages
Convenience sampling	Recruit students from an undergraduate participant pool, or ask students around you at lunch or in class.	Inexpensive, efficient, convenient.	Likely to introduce bias into the sample. Results may not generalize to intended population.
Purposive sampling	In a convenience sample, select people who meet a criterion (e.g., an age group).	Relatively convenient. Sample includes only types of people you are interested in.	Likely to introduce bias into the sample. Results may not generalize to intended population.
Quota sampling	Identify the proportion of each important subgroup within a population, and then use convenience sampling within each subgroup.	Inexpensive, efficient, convenient. Slightly more sophisticated than convenience sampling.	Likely to introduce bias into the sample. Results may not generalize to intended population. No method for choosing individuals in subgroups.

Probability Sampling

Simple Random Sampling

With *simple random sampling*, every member of the population has an equal probability of being selected for the sample. If the population has 1,000 members, each has one chance out of a thousand of being selected. Suppose you want to sample the students who attend your school. Based on a list of all students enrolled, you could choose a sample at random with every person on the list having an equal chance of being included. Whenever people are randomly selected from a specific population to participate in a study, the resulting sample is called a *random sample*.

Stratified Random Sampling

A somewhat more complicated procedure is *stratified random sampling*. In this technique, the population is divided into subgroups (or strata), and then simple random sampling is used to select sample members from each subgroup. Any number of dimensions could be used to divide the population, but the dimension(s) chosen should be relevant to the problem under study. For instance, a survey of sexual attitudes might stratify on the basis of age, sex, and sexual orientation, because these factors are related to sexual attitudes. Random sampling would then be employed within each subgroup to maximize the representativeness of the sample, along the dimensions specified.

One advantage of stratified random sampling is that the sample will accurately reflect the composition of the various subgroups. This kind of accuracy is particularly important when some subgroups represent very small percentages of the population. For example, a Canada-wide survey might stratify its sample on the basis of province, age, and sex, to best reflect these population characteristics. If Indigenous people make up 5 percent of a city of 100,000, a simple random sample of 100 people might not include any Indigenous people at all. In contrast, a stratified random sample divided by ethnicity would include five Indigenous people chosen randomly from the Indigenous population. In practice, when it is important to represent a small group within a population, researchers will typically “oversample” that group to ensure that a representative sample of the group is surveyed. This is because a large-enough sample must be obtained to be able to make inferences about that sub-population.

Cluster Sampling

It might have occurred to you that obtaining a list of all members of a population might be difficult. What if officials at your school decide that you cannot have access to a list of all students? What if you want to study a population that has no list of members, such as people who work in regional health care agencies? In such situations, a technique called *cluster sampling* can be used. Rather than randomly sampling from a list of people, the researcher can identify “clusters” of people and then sample from these clusters. After the clusters are chosen, all people in each cluster are included in the sample. For example, you might conduct the survey of students using cluster sampling by identifying all classes being taught—the classes are the clusters of students. You could then randomly sample a subset based on this list of classes.

Non-probability Sampling

In contrast, non-probability or *non-random* sampling techniques are quite arbitrary. A population may be defined, but little effort is expended to ensure that the sample accurately represents the population. As a result, what we find in the sample may not generalize to the population. That said, non-probability samples are cheap and convenient, and as a result quite prevalent in the behavioural sciences. Three types of non-probability sampling are convenience sampling, purposive sampling, and quota sampling.

Convenience Sampling

Perhaps the most prevalent form of non-probability sampling is *convenience sampling* (sometimes called *haphazard sampling*). In this form of non-probability sampling, participants are recruited wherever you can find them. To select a sample of students from your school, you might stand in front of the student union building at 9 a.m. and ask passers-by, or ask people who sit around you in your classes to participate, or perhaps visit a couple of on-campus residences. Unfortunately, such procedures are likely to introduce biases into the sample, such that the sample may not accurately

represent the population of all students. For example, if you selected your sample from students walking by the student centre at 9 a.m., your sample excludes students who don't frequent this location, and it may also eliminate afternoon and evening students. This might mean that your sample could be younger and work fewer hours at paid employment than the general population of students. Sample biases such as these limit your ability to use your sample data to estimate actual population values. Your results may not generalize to your intended population, but instead may describe only the biased sample that you obtained.

Purposive Sampling

A second form of non-probability sampling is *purposive sampling*. This refers to sampling for a specific purpose to obtain a sample of people who meet some predetermined criterion. For example, imagine researchers using purposive sampling while asking cinema customers to fill out a questionnaire. Instead of sampling anyone walking toward the theatre, they look at each person to make sure that they fit some criterion—under the age of 30 or an adult with one or more children, for example. This is one way to limit your sample to a certain group of people. However, it is still not a probability sample, which limits the generalizability of results based on this sample.

Quota Sampling

A third form of non-probability sampling is *quota sampling*. A researcher who uses this technique chooses a sample that reflects the numerical composition of various subgroups in the population. Thus, quota sampling is similar to the stratified sampling procedure previously described, but without the randomness. To illustrate, suppose you want to ensure that your sample of students includes 19 percent first years, 23 percent second years, 26 percent third years, 22 percent fourth years, and 10 percent graduate students because these percentages reflect the distribution in your school's total population. A quota-sampling technique would make sure you have these percentages, but you would still collect your data using convenience techniques. If you didn't get enough graduate students in front of the student union, perhaps you could go to a graduate class to complete the

sample. Although quota sampling is a bit more sophisticated than convenience sampling, the problem remains that no restrictions are placed on how people in the various subgroups are chosen. The sample does reflect the numerical composition of the whole population of interest, but respondents within each subgroup are selected in a haphazard manner, and not everyone in the population has a known or equal chance of being included.

Page 146

Reasons for Using Convenience Samples

Much research in psychology uses non-probability sampling techniques to obtain participants for surveys, experiments, and other studies. The advantage of these techniques is that the investigator can obtain research participants without spending a great deal of money or time to select the sample. For example, it is common practice to invite students in introductory psychology classes to participate in studies being conducted by faculty at their institution.

Even in studies that do not use university students, the sample is often based on convenience rather than concern for obtaining a random sample. It is common for researchers who study children to draw from one or two particular elementary schools with whom they have a good relationship. This method introduces bias because only children from particular neighbourhoods with certain social and economic characteristics are included, and so the results may not generalize to all children.

Why aren't researchers more worried about obtaining random samples from the "general population" for their research? Some researchers have argued that basic psychological research studies phenomena that should operate the same way in all humans. What is discovered in university undergraduates is likely to be true of all other humans on earth, especially for basic phenomena like perception. However, critics have pointed out that this does not appear to be true, with the results observed among people from Western, educated, industrialized, rich, and democratic societies differs from those observed in other populations, even for very basic processes (Henrich, Heine, & Norenzayan, 2010a, 2010b). This is a topic we return to in [Chapter 14](#).

That said, convenience samples are not totally uninteresting. In fact, university undergraduates might be very interesting! It is just important to specify to what populations we believe our samples generalize, and justify our reasons for believing this. In addition, results based on a non-representative non-probability sample might lead other researchers working with other populations to try and replicate the result. A single study isn't expected to be without flaws or to incorporate samples from all populations of interest. Rather, science is the process of accumulating over time. If future studies conducted in other populations, in other cultures for example, also find a similar result, then we might come to learn more about the generalizability of a finding. Replicating results with multiple samples and multiple methods is one way to increase the external validity of research, while using non-probability sampling ([Chapter 14](#)). The results of many studies can then be synthesized to gain greater insight into the phenomenon being investigated (cf., Albright & Malloy, 2000).

Keep in mind that the generalizability of non-probability samples will depend in part on the population of interest. Introductory psychology students at your school are probably quite representative of the first-year students at your school in general. These introductory psychology students are probably a little bit less representative of all the students in your school (e.g., first-year students are somewhat different from fourth-year students), even less representative of all university students in your province, even less so for all university students in the nation, and even less so for all Canadians (regardless of age), and so forth. We will revisit this discussion about the *external validity* of our samples in [Chapter 14](#).Page 147

We have considered many issues that are important when asking people about themselves. If you engage in this type of research, you may need to design your own questions by following the guidelines described in this chapter and consulting other sources (e.g., Judd et al., 1991; Gosling & Johnson, 2010). You can also adapt questions and entire questionnaires that have been used in previous research. For example, Greenfield (1999) studied Internet addiction by adapting questions from a large body of existing research on addiction to gambling. Consider using previously developed questions, particularly if they have been useful in other studies (make sure you don't violate copyrights, however). For examples, see

compilations of various measures of social and political attitudes developed by others (e.g., Robinson, Rusk, & Head, 1968; Robinson, Shaver, & Wrightsman, 1991; Registry of Scales and Measures, www.scalesandmeasures.net; your library may also have access to the APA's database, *PsycTESTS*).

We noted in [Chapter 4](#) that both non-experimental and experimental research methods are necessary to fully understand behaviour. [Chapters 6](#) and [7](#) have focused on techniques that commonly form the backbone of many non-experimental designs, yet are also useful in operationalizing dependent variables in experiments. In the next chapter, we begin to explore the design of experiments in detail.

Study Terms

Test yourself! Define and generate an example of each of these key terms.

- *closed-ended questions* (p. 132)
- *cluster sampling* (p. 144)
- *confidence interval* (p. 140)
- *convenience sampling* (p. 145)
- *external validity* (p. 141)
- *focus group* (p. 139)
- *graphic rating scale* (p. 134)
- *inattentive responding* (p. 131)
- *interviewer bias* (p. 138)
- *non-probability sampling* (p. 143)
- *open-ended questions* (p. 132)
- *panel study* (p. 126)
- *population* (p. 140)
- *probability sampling* (p. 143)
- *purposive sampling* (p. 145)
- *quota sampling* (p. 145)
- *random sample* (p. 143)

- rating scale (p. 132).
- response rate (p. 142).
- response set (p. 127).
- sampling (p. 140).
- sampling error (p. 140).
- sampling frame (p. 142).
- semantic differential scale (p. 134).
- simple random sampling (p. 143).
- stratified random sampling (p. 143).
- survey research (p. 125).
- “yea-saying” or “nay-saying” response sets (p. 131).

Review Questions

Test yourself on this chapter's learning objectives. Can you answer each of these questions?

1. What is a survey? List some research questions you might address with a survey.
2. What are some features to consider when constructing questions for surveys (including both questions and response alternatives)?Page 148
3. What is a social desirability response set? What can a researcher do to identify and/or minimize it?
4. Compare the different ways to administer a survey using questionnaires and interviews. What are the advantages and disadvantages of each method?
5. How do sample size, sampling frame, response rate, and sampling technique affect the interpretation of survey results?
6. Distinguish between probability and non-probability sampling techniques. When would a researcher use each of these techniques? What are the costs and benefits involved in each technique?
7. Compare and contrast convenience sampling, purposive sampling, and quota sampling.
8. Compare and contrast simple random sampling, stratified random sampling, and cluster sampling.

Deepen Your Understanding

Develop your mastery of these concepts by considering these application questions. Compare your responses with those from other people in your study group.

1. In the Dumont et al. (2009) study on teenage employment (discussed [above](#)), longer work hours were associated with lower grade point averages. Can you conclude that working longer hours *causes* lower grades? Why or why not? How might you expand the scope of this investigation using a panel study? An experiment?
2. Select a topic for a survey. Write at least five closed-ended questions that you might include. For each question, write one “good” version and one “poor” version. For each poor question, state what elements make it poor and why the good version is an improvement.
3. Suppose you want to know how many books in a bookstore have only male authors, only female authors, or both male and female authors. (You might operationally define the “bookstore” as a large retail store, the textbook section of your university bookstore, or all books in the stacks of your library.) Because there are thousands of books in the store, you decide to study a sample rather than examine every book there. Describe how you might sample books using a non-probability sampling technique. Then describe a possible probability sampling technique. How might your results differ using the two techniques?

Experimental Design



©MayaCom/Getty Images

The essence of experimental design is exposing participants to two different conditions that are identical except for the independent variable of interest, which you've manipulated. Like playing with one of two nearly identical alpacas, except one is brown and one is white.

Learning Objectives

Keep these learning objectives in mind as you read to help you identify the most critical information in this chapter.

By the end of this chapter, you should be able to:

1. [LO1](#) Describe how confounding variables affect the internal validity of an experiment.
2. [LO2](#) List and explain three major steps toward planning a basic experiment.
3. [LO3](#) Describe the pretest-posttest design, including the advantages and disadvantages of using a pretest.
4. [LO4](#) Describe the matched-pairs design, including reasons to use this design.
5. [LO5](#) Contrast a between-subjects design with a within-subjects design, including advantages and disadvantages of each.

Page 150The experimental method is a valuable tool in the researcher's toolbox. The next four chapters on experimentation build on the basic concepts you first learned in [Chapter 4](#) (e.g., independent variable, dependent variable, random assignment). Suppose you want to test the hypothesis that crowding impairs cognitive performance. To do this, you might put one group of people in a crowded room and another group in an uncrowded room, then have both groups complete a cognitive test. In this case, crowdedness is the independent variable, and crowded versus uncrowded are two levels or conditions of this independent variable. The cognitive test that both groups complete is the dependent variable (with the same test used for both groups). Now, suppose that the people in the crowded condition do not perform as well on the cognitive tests as those in the uncrowded condition. Can the difference in test scores be attributed to the difference in crowding? The answer is yes, but only if the experiment is well-designed and properly conducted. For example, if there is no other difference between the conditions besides the level of crowdedness, then we would feel more confident in inferring that crowdedness caused this difference in test performance between groups. But if the conditions differed in any other way, then we would have difficulty knowing what caused the difference in performance between groups. Imagine if participants in the crowded condition were tested in a room with nine other participants, but those in the uncrowded condition were tested in a room with just one other participant. In this case, both the amount of crowding and the number of people in the room differs between condition. As a result, it is impossible to know whether the difference in scores between conditions is due to feeling crowded or the number of other people in the room. How might you avoid this problem?



LO1 Confounding and Internal Validity

Recall from [Chapter 4](#) that a critical advantage of experiments over non-experiments is that experiments can support inferences of causality (when conducted properly). From experiments, we can learn that altering one variable *causes* another variable to change. The researcher manipulates the independent variable, creating groups that differ in the levels of that variable, and then examines the effect of this manipulation on the dependent variable. Importantly, all other variables must be kept constant through various means including direct experimental control and random assignment (or an equivalent procedure). If the scores for the dependent variable are different between the groups (the levels of the independent variable), the researcher concludes that the independent variable caused this difference in results. This conclusion is based on the logic that if the only difference between the groups is the manipulated independent variable, then this difference must be the cause of any difference in the measured dependent variable.

Although the logic of an experiment is quite simple, in practice there are numerous possible pitfalls when designing an experiment. If a researcher happens to fall into one of these pitfalls, the entire logic of the experiment

will be undermined and we will not be able to make any causal inferences. Consider the example of the crowding experiment, described above. Both group size and crowding change between conditions, and both might influence cognitive performance. Because we are interested in crowding, group size becomes a potential confounding variable. Recall from [Chapter 4](#) that a confounding variable is a variable not of interest to the researcher that varies (i.e., changes) along with the independent variable, and could provide an explanation for the results observed. In this case, the researchers want to study crowding, not the number of people in the room, per se: So the number of people in the room is not a variable of interest. However, in the design proposed above, as crowding increases so does the number of people in the room. The independent variable (i.e., crowding) and an uncontrolled variable of no interest (i.e., number of people) are completely intertwined and their influences cannot be separated. Lastly, the number of people could also explain why those in the crowded condition did worse: Perhaps they were just more distracted by the presence of other people than those in the uncrowded condition (who only had one other person in the room). So, how do we avoid this confound? One way might be to test all participants in groups of ten, and vary crowding by the seating arrangement (e.g., spread apart or tightly packed). Now we have controlled the number of people in the two conditions: Both the crowded and uncrowded condition have exactly ten people in the room.

☆ Student Spotlight: Anticipating Confounds ☆

Anticipating confounds is often very difficult. It takes a lot of careful thought and analysis to identify confounds and then control for them. Maria Kotovych and Mark Holden, working with Drs. Peter Dixon and Maria Bortolussi at the University of Alberta, conducted a series of studies about identifying with fictional characters. Specifically, they wondered whether first-person narratives (i.e., with a narrator using “I” or “we”) help readers make sense of characters, because readers use their own experiences to clarify the thoughts and actions of characters. So they conducted an experiment in which a story was altered in various ways, including adding information that a reader would normally infer (i.e., the manipulation of the independent variable). After completing their first experiment, they realized that there may have been a potential confound, since the material they

added was written in quite a different style from the original text. So in a follow-up study, they created two versions of the original story that were both manipulated by adding text in the same style, but these changes were either directed at things readers were expected to infer (i.e., as in the first study) or unrelated content. This clever follow-up allowed them to uncover whether their first set of results were due to a confound or not (i.e., the presence of text in a different style). A third experiment was also conducted, and this package of studies was published in the journal *Scientific Studies of Literature* (Kotovych, Dixon, Bortolussi, & Holden, 2011).

Good experimental design allows researchers to make causal claims because competing, alternative explanations (e.g., confounding variables) are controlled or eliminated. When the results of an experiment can confidently be attributed to the independent variable, the experiment is said to have high internal validity ([Chapter 4](#)). To achieve high internal validity, the researcher must design and conduct the experiment so that only the independent variable can be the cause of the results. However, identifying and eliminating confounds and other threats to internal validity is often very difficult. We will revisit this idea of threats to internal validity in [Chapter 10](#) in the context of quasi-experimental designs, which are used when random assignment (or an equivalent procedure; see [below](#)) is impossible. For now, let us continue considering essential features of experiments.

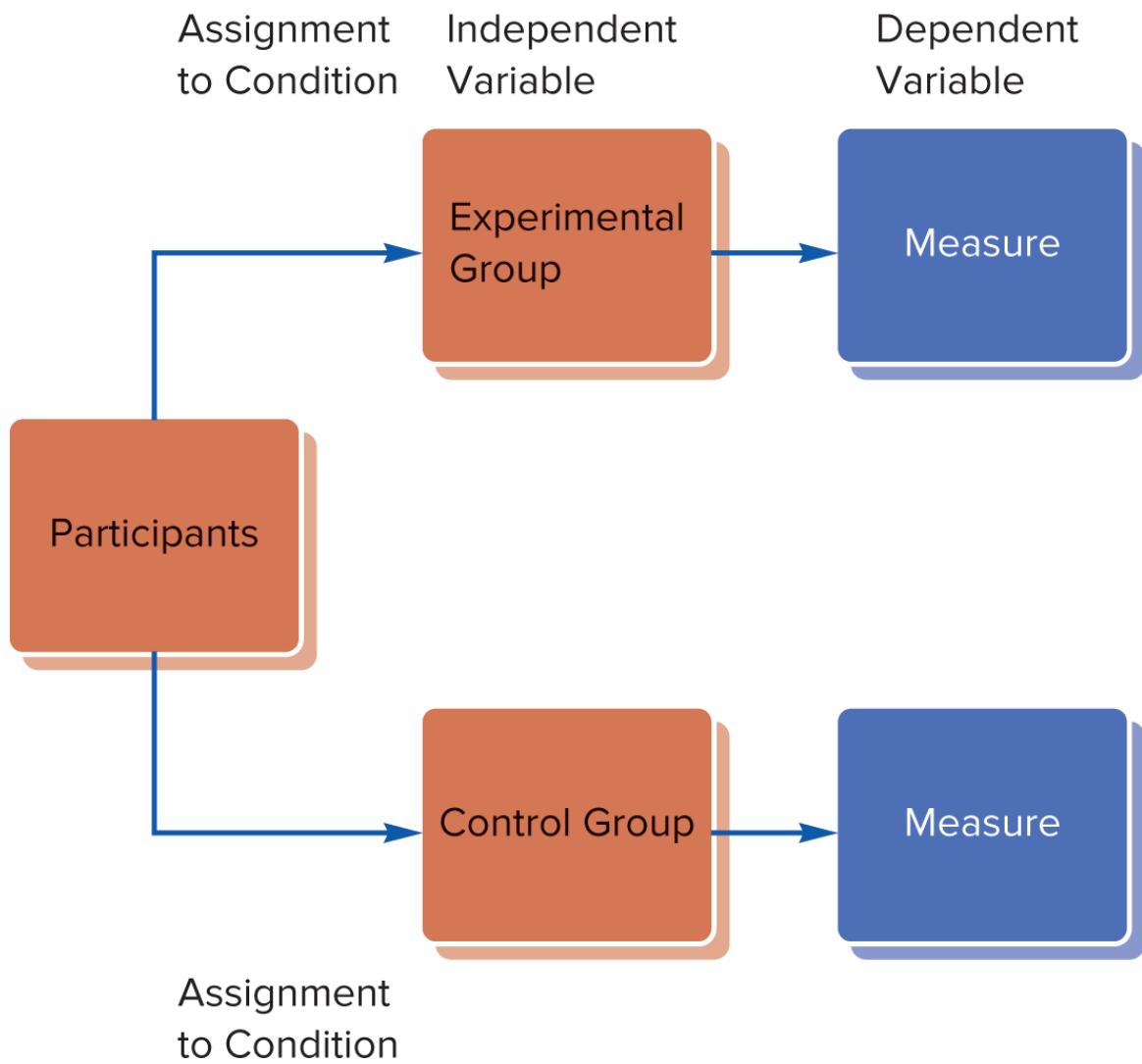


LO2 Planning a Basic Experiment

The simplest possible experimental design has two variables: the independent variable, which itself has two levels or conditions (e.g., an experimental group and a control group), and the dependent variable. (In [Chapter 11](#), we will explore more complex experimental designs with more levels of an independent variable or more than one independent variable.) Researchers must make every effort to ensure that the only difference between the two levels of the independent variable—the experimental group and the control group—is the manipulated variable. There are also two broad classes of experiment. The first is a [between-subjects design](#) (also known as an *independent groups design*), in which different people experience the levels of the independent variable. In this kind of experiment, the experimental and control groups are distributed between different people. The second is a [within-subjects design](#) (also known as a *repeated measures design*), in which all the same people experience all levels of the independent variable. For a within-subjects design, the experimental and control conditions are equally distributed within all participants. We will begin by discussing a between-subjects design, then move to within-subject designs.

Any between-subjects experiment involves three broad steps: (1) obtaining two approximately equivalent groups of participants, (2) introducing different levels of the independent variable to the participants, and (3) measuring the dependent

variable. Steps 2 and 3 also hold true for within-subject designs. When there are only two levels of the independent variable, the between-subjects experimental design can be diagrammed like this:



The first step in planning a between-subjects experiment is to decide how to assign participants to the levels of the independent variable. The procedures used must create equivalent groups and must eliminate any potential selection differences: The people selected to be in the experimental condition should not differ in any systematic way from those selected for the control condition. For example, if mostly high-income participants are assigned to the experimental condition, and mostly low-income participants to the control condition, this would result in a selection difference; now income is confounded with the

independent variable. For a between-subjects experiment, researchers try to ensure that the participants in each condition are approximately equivalent using either random assignment or a matched-pairs design. In the matched-pairs design, groups are made equivalent by first selecting pairs of participants who score the same (i.e., are matched) on some variable of interest, and then using random assignment to determine which person in each pair will experience which condition. For within-subject experiments, the groups of participants for each level of the independent variable (i.e., each condition) are already equivalent by virtue of the design. We will examine each of these designs in detail throughout this chapter.

Page 153



Think about It!

Why is there no need to make groups equivalent, through random assignment or some other means, when using a within-subjects experimental design? How can we be sure that these groups are, in fact, similar using this kind of experiment?

The second step in planning any basic experiment, be it between-subjects or within-subjects, is to operationalize (or operationally define) the independent variable (see [Chapter 4](#)), creating at least two levels. Sometimes, these levels include an experimental group that receives a treatment and a control group that does not. For example, a researcher might study the effect of reward on motivation by offering a reward to one group of children before they play a game and offering no reward to children in the control group. For studies examining whether a treatment intervention works, one condition will likely involve administering the intervention, with the other condition being a control group that did not receive this intervention. Other times, researchers may use two different amounts of the independent variable. This could be more reward in one group than the other, or more hours of treatment to one group compared to the other. Any of these approaches would provide a basis for comparing the two groups.

The third step in planning an experiment is to operationalize the dependent variable, which will allow us to measure the effect of the independent variable on the dependent variable. The same measurement procedure is used for both conditions, so that they can be compared. When the groups are equivalent and there are no confounding variables, or other threats to internal validity, we can conclude that difference between conditions for scores on the dependent variable are caused by the independent variable. A statistical significance test would

typically be used to assess the difference between the groups (see [Chapter 13](#)). Importantly, an experiment must be well-designed, and confounding variables must be eliminated, before we can draw appropriate conclusions from statistical analyses.

Between-Subjects Experiments

In a between-subjects design (also called an independent groups design), different participants are assigned to each condition, or level of the independent variable. In order to ensure that the groups experiencing each condition are roughly equivalent, who ends up in each group is typically determined using random assignment. This means that assigning people to the different conditions is determined by chance. Researchers can use a list randomizer like the one available at www.random.org to assign participants to each condition ([Chapter 4](#)), or some other method of producing a random order (e.g., generating random numbers in a spreadsheet and then sorting participants by these). Random assignment ensures that the groups will be approximately equivalent in terms of a whole host of participant characteristics, such as income, intelligence, age, or political attitudes. By allowing chance to determine who experiences each condition, those who are older have an equivalent chance of ending up on the control condition as those who are younger. With a large number of participants, random assignment increases the likelihood that nuisance variables (variables not of interest) related to participant characteristics will be approximately equally distributed across conditions. In this way, participant characteristics cannot be an alternative explanation for the experiment's results. It is impossible for different scores on the dependent variable between conditions to be attributed to participant characteristics, when these characteristics are roughly equivalent between the groups. How many participants you need in each group, both for random assignment to establish group equivalence and to detect effects, is a complicated question with many factors to consider. It is generally safest to collect as many people as you possibly can, to maximize the effects of random assignment, to detect smaller effects, and to generalize to the population from which you've drawn your sample. As a starting point, consider 50 people per condition for a simple two-condition between-subjects experiment, a bare minimum (Simmons, Nelson, & Simonsohn, 2013). But you should aim for much larger samples, ideally.

Page 154



LO3 Pretest-Posttest Design

Sometimes a researcher needs to be extra cautious that random assignment to condition created groups that were equivalent on some particular variable (e.g., trait extraversion). Therefore, the researcher may choose to add a pretest to measure that variable before any experimental manipulation. Pretest scores for the two groups are then compared to ensure that the two groups were approximately equivalent on the critical variable, before the manipulation was introduced. In many cases, a pretest will be used to measure levels of the key dependent variable, to ensure that participants in both conditions had approximately equal levels of the construct presumed to be affected by the independent variable and to more accurately measure change in this construct. This design is sometimes called a *pretest-posttest design*. When no pretest is given for a between-subjects design, this is sometimes called a *posttest-only design*. The pretest-posttest design makes it possible for researchers to be absolutely sure that the groups were equivalent at the beginning of the experiment for a crucial variable. In this terminology, the posttest refers to the dependent variable, measured after the experimental manipulation of the independent variable.

Advantages and Disadvantages of a Pretest

The pretest is a precaution that is usually not necessary if participants have been randomly assigned to the two groups. However, there are three main reasons why a researcher may add a pretest: (1) to counter problems

associated with a small sample size, (2) to select appropriate participants, and (3) when participants might drop out of the study. First, although random assignment is likely to produce equivalent groups, as sample size decreases, it becomes less likely that the groups will be approximately equal. Thus, a pretest enables the researcher to assess whether the groups were already roughly equivalent on some critical variable before the manipulation began.

Second, a pretest may be used to select the participants to include in the experiment. A researcher might need to give a pretest to find the lowest or highest scorers on a measure of smoking, math anxiety, or prejudice, for example. Once the target participants are identified, they would be randomly assigned to the experimental or control group. However, this selection method may introduce another problem: regression toward the mean ([Chapter 10](#)), which is another threat to internal validity for an experiment.

Third, a pretest may be useful when there is a good possibility that participants will drop out of the experiment, in order to ensure that those who remain do not differ across conditions (producing a selection bias). Participants choosing to leave a study can produce a problem known as [selective attrition](#), with participant dropout most likely to occur in studies that last over a period of time (i.e., longitudinal studies). When participants leaving a study produces a difference between conditions, this selective attrition is a threat to internal validity ([Chapter 10](#)). People may drop out for reasons unrelated to the experimental manipulation, such as illness, and this may not result in differences between conditions. However, if people are dropping out for some reason related to the manipulation, this selective attrition may cause a difference between groups (i.e., a selection bias). Even if the groups are equivalent at the beginning, different rates of dropout can make them non-equivalent by the time the dependent variable is measured at the end of the study. Imagine a study of a treatment to reduce smoking. One possibility is that the heaviest smokers, those most addicted, in the experimental condition (i.e., treatment intervention) might be more likely to leave than those in the control condition (i.e., no treatment). Therefore, when the posttest is given, only the light smokers would remain in the experimental condition, so a comparison between

groups would show less smoking in the experimental group, even if the program had no effect. In this way, selective attrition becomes an alternative explanation for the results. Pretest scores enable researchers to examine whether selective attrition is a plausible alternative explanation for differences between groups, by examining the scores on the crucial variable for those who left and those who remain.^{Page 155}

Thus, pretests offer some advantages for a between-subjects experiment. One disadvantage of a pretest, however, is that a pretest can sensitize participants to what you are studying, enabling them to figure out your hypothesis (i.e., it may create a demand characteristic, another threat to internal validity; [Chapter 9](#)). They may then react differently to the manipulation than they would have without the pretest. When a pretest influences the way participants react to a manipulation, it is very difficult to generalize these results to people who have not received a pretest; that is, in these cases, the independent variable may not have an effect in the real world, where pretests are rarely given ([Chapter 14](#)).

For researchers who decide a pretest is useful, there are a few ways to avoid or measure its potentially detrimental influence on the results for a study. If awareness of the pretest is a problem, the pretest can be disguised using deception ([Chapter 3](#)). One way to do this is to embed the pretest in a set of irrelevant measures, so that it is not obvious that the researcher is interested in a particular topic. ([Chapter 5](#) has additional tips for dealing with reactivity.) Another way is to examine the influence of a pretest directly, by including the presence or absence of a pretest as another condition, embedding the levels of the independent variable within those two groups (known as a Solomon four-group design; for an example, see Wertz Garvin & Damson, 2008).



LO4 Matched Pairs Design

Another way to make groups approximately equivalent for a between-subjects experiment is to use a *matched pairs design*, which is sometimes known as a yoked design. (A yoke is a piece of wood that connects two animals, such as two oxen, working in a pair.) Instead of simply randomly assigning participants to groups, as is typical for a between-subjects experiment, a matched pairs design begins with matching people on a crucial participant characteristic. This matching variable can be either the dependent variable itself or another variable that is strongly related to it. For example, a researcher interested in studying the effect of aging on cognitive functioning might start by pairing participants who are close in age, then randomly assigning one person in the pair to the experimental condition and the other to the control condition. In this way, the experimental and control conditions become closely matched in terms of ages.

As an example of a matched pairs design, Hockey and Earle (2006) looked at whether having control over one's work schedule influenced how tired people felt compared to being told what one's schedule will be. In the first step of matching, all participants were measured on some variables related to this topic (e.g., need for control, typing speed). Participants were then matched to create pairs of people who were similar to each other on all of these variables. All participants were subsequently given numerous office tasks to complete (e.g., typing, scheduling). Crucially, one member of each

pair was randomly assigned to have control over their own schedule, with the matched participant told to complete the tasks in the same order and duration as what their match had chosen. In this way, matching was used to achieve equivalency of the groups, reducing the chance of selection bias. What they found was that when people control their own work schedules, they report feeling less tired after working than people who are told exactly what to do. Page 156

Matching is most likely to be used when it is not possible to collect a large sample, making simple random assignment less likely to be successful in creating equivalent groups. Sometimes only a few participants are available or it is too costly to collect a large number of participants. The matched pairs design helps to remove some variability between groups related to participant characteristics. Therefore, the matched pairs design may be particularly useful when individual differences are expected to have a large influence on the dependent variable. However, matching procedures can be costly and time-consuming, depending on the number of characteristics used for matching. Sometimes these efforts are worthwhile only when you believe that the matching variables are strongly related to the dependent measure.

Within-Subjects Experiments

Instead of creating two separate groups and distributing the levels of the independent variable between these groups, an alternative procedure is to distribute conditions equally within all participants. In this design, a within-subjects design or repeated measures design, all participants experience all conditions (e.g., both the experimental and control condition). This ensures that the groups for each condition are absolutely identical, removing all possibility of a selection bias or differences in participant characteristics between groups. In this kind of experiment, participants are measured on the dependent variable after being in each condition of the experiment. Pretests can also be used in a within-subjects design, providing the advantages mentioned [earlier](#).

Let's consider an example. Have you ever thought about listening to audio books rather than reading? How often do you read out loud? University of Waterloo researchers used a within-subjects design to test whether these different ways of consuming information influence people's memory for the material and their tendency to mind-wander (Varao Sousa, Carriere, & Smilek, 2013). Passages were taken from the same book, and the independent variable was the way participants encountered them, including three levels or conditions: (1) reading silently to themselves, (2) reading aloud, and (3) listening to them being read by someone else. While reading the passages, all participants were interrupted regularly and asked to report whether their minds were wandering. Afterward, they completed a short memory test for the material. This procedure was repeated for the other two conditions, until all participants had completed all three conditions. The least mind-wandering and greatest memory was found for material that people read aloud, followed by material that was read silently, with material encountered by listening coming last (regardless of the order they were administered). Based on these findings, you should consider using audio books only for material you don't need to remember!



LO5 Advantages and Disadvantages of the Within-Subjects Design

The within-subjects design has several advantages, as well as disadvantages, when compared with between-subjects designs. A major advantage is that fewer research participants are needed because everyone participates in all conditions. Imagine you need 50 people per condition for an experiment: For a two-condition between-subjects design, this means 100 people in total, versus just 50 people for a within-subjects version of the same experiment. When participants are scarce, or when it is costly to run each participant, a within-subjects design can maximize the amount of data collected. Consider the importance of these issues for research relying on specialized machinery resulting in costs of hundreds of dollars per participant (e.g., fMRI), or unique populations that take years of searching to find just a few people who meet the criteria (e.g., people with rare conditions).Page 157

An additional advantage of within-subjects designs is that they are extremely sensitive, able to detect small differences between conditions. When the exact same people are in both conditions, less of the variance in the data is attributed to error (e.g., error associated with differences between the people in each condition). To illustrate what this means, consider hypothetical data from the reading experiment described above. Using a between-subjects design, the first three participants in the read silently condition had memory scores of 58, 71, and 82. In contrast, the first three participants in the listen condition had scores of 54, 68, and 75. This means that the average memory score was higher when people read silently (average of 70.33) rather than listened (average of 65.66). However, there is a lot of variability in the scores in both groups: The scores are quite

different from one another. Not everyone in the read silently condition had good memory and not everyone in the listen condition had poor memory. A major reason for this variability is that people differ. There are individual differences in memory, so there is a range of scores in both conditions. Because we are not interested in these individual differences, for the purposes of this particular study, this variability can be characterized as part of the *error* in the scores (note that it's not wrong or bad, just not of interest). We cannot explain this variability, and it affects our statistical analyses ([Chapter 13](#)). Just like when we considered reliability of measures ([Chapter 5](#)), we want as little error as possible.

However, if the same scores were obtained from the first three participants in a within-subjects design, the conclusions would be much different. Let's line up the hypothetical memory scores for the two conditions:

	Listen	Read Silently	Difference
Participant 1	54	58	+4
Participant 2	68	71	+3
Participant 3	75	82	+7

Regardless of condition, Participant 1 seems to have worse memory overall than Participant 3; these overall differences linked to specific individuals are what we mean by individual differences in memory. In a within-subjects design, we have measured these systematic individual differences and separated them from random error and from the effect of the independent variable. The *difference* column shows us that scores are higher for every participant in the read silently condition compared to the listening condition. Note that not every difference score is the same, so there is still some unexplained random error. Nonetheless, by accounting for individual differences in a within-subjects design, we are better able to detect an effect of the independent variable on the dependent variable, if one exists.

The major challenge with a within-subjects design stems from the fact that the different conditions must be presented in a particular order. This problem doesn't exist for a between-subjects design, because in that design every participant only experiences one of the conditions. Order can become a confound if left unaddressed. Imagine, hypothetically, that the read aloud condition always came third, and participants remembered this material best relative to all other conditions. Although this result could be caused by the manipulation of the independent variable (i.e., reading aloud versus listening or reading), the result could also simply be an [*order effect*](#): Perhaps the order of the conditions affects

the dependent variable. In this case, maybe memory is best for the third condition, simply because of the practice gained from doing the earlier tasks.

There are several types of order effects, which all pose a threat to the internal validity of a study. Time-related order effects are possible whenever there is a sequence of tasks to perform: These include practice effects and fatigue effects. A *practice effect* occurs when performance improves because of repeated practice with a task. In the example above, perhaps people became more accustomed to the memory tests and developed better strategies for remembering information, and this is why performance is best for the third condition. A *fatigue effect* occurs when performance worsens as participants become tired, bored, or distracted. This is just like the fatigue effect we discussed with respect to long surveys. If the listening condition was always presented third, and participants recalled the least while listening, an alternative explanation for these worse memory scores could be fatigue or boredom. Page 158

Other types of order effects occur when the effect of the first condition carries over to influence the way people respond to the second condition. A *contrast effect* occurs when the response to the second condition in the experiment is altered because experiences of the first highlight how they are different. Suppose the independent variable in an experiment is the severity of a crime, with the dependent variable being assigned punishment. After reading about murder in the first condition, reading about theft in the second condition might make this crime of theft seem much milder to participants than it would have seemed if it were presented first (or on its own). People might punish the thief less harshly when it follows a description of murder, as a result. In this case, the order of the conditions is confounded with the severity of the crime described.

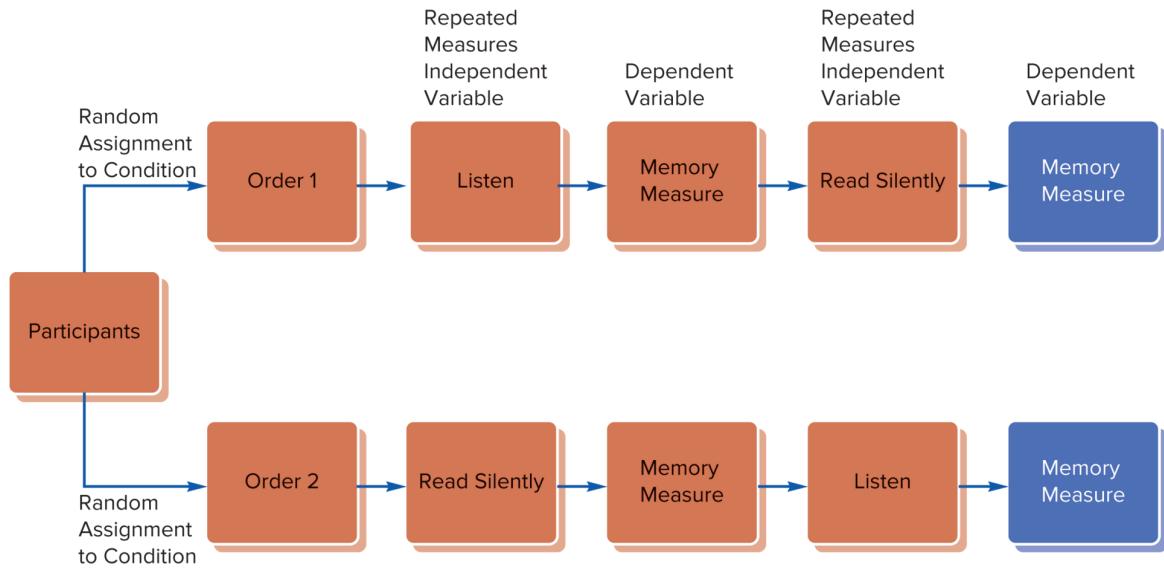
There are two ways to deal with order effects in within-subjects designs: (1) using counterbalancing techniques and (2) ensuring that the time between conditions is long enough to minimize the influence of the first condition on the second. These strategies can also be combined. We consider each in turn.

Counterbalancing

Complete Counterbalancing

In a within-subjects design that uses complete *counterbalancing*, all possible orders of presentation for the conditions are included in the experiment. Let us consider counterbalancing for just two of the conditions in the memory study

discussed earlier: listening and reading (Varao Sousa et al., 2013). To include all possible orders of presentation for the two conditions, half of the participants would be randomly assigned to the “listen then read” order, and the other half to the “read then listen” order. This design is illustrated as follows:



By counterbalancing the order of conditions, any effect of order is equally distributed between the two conditions. In other words, if practice effects are boosting scores for the condition presented second, this is not a problem as both conditions appear second for exactly half of the participants. Similarly, if fatigue effects are lowering scores for the condition presented later, that's not a problem as this occurs equally across the group (for half the participants, the second condition is reading, and for the other half, it is listening).

Counterbalancing does not remove the potential influence of order effects, but it ensures that these influences are spread out equally among the conditions. This way, any difference between conditions cannot be due to an effect of order: Order is influencing both conditions equally. In addition, this design allows us to actually examine whether any order effects are occurring. We can compare average memory scores for the listen condition when it comes first versus second. If they are the same, then there is no evidence of any order effects. Similarly, we can detect whether a fatigue effect might be operating, because memory will always be worse for the second condition, regardless of whether that is the listen or read condition.

Counterbalancing principles can also be extended to experiments with three or more levels of the independent variable. With three levels, there are six possible orders (calculated as the factorial of 3 ($3! = 3 \times 2 \times 1 = 6$)). With four levels, the number of possible orders increases to 24 ($4! = 4 \times 3 \times 2 \times 1 = 24$). The number of possible orders increases quickly as more conditions are added. In an experiment with 10 conditions, there are 3,628,800 possible orders! Fortunately, there are alternatives to complete counterbalancing that still allow researchers to draw valid conclusions about the effect of the independent variable.

Partial Counterbalancing

One technique to control for most order effects without having to run all possible orders is to employ a [Latin square design](#). The Latin square design uses a limited set of all possible orders carefully constructed to ensure that (1) each condition appears first, second, third, and so on; and that (2) each condition appears directly before and directly after each other condition exactly once. Consider an experiment in which participants' memory and reaction time were measured after varying amounts of sleep the previous night: eight, six, four, and zero hours of sleep (Roehrs, Burduvali, Bonahoom, Drake, & Roth, 2003). Instead of having 24 different orders in this within-subjects design, the researchers used a Latin square. The number of orders in a Latin square is equal to the number of conditions, so there were four orders, as depicted in [Table 8.1](#). For more details on the Latin square, there is a seminal paper by Grant (1948) and also many resources online that will help to create a Latin square for you based on your study parameters. More details are also provided in [Appendix E](#).

Table 8.1 A Latin Square with Four Order Conditions, Based on Roehrs et al. (2003)

Ordinal Position of Within-Subjects Conditions

	First	Second	Third	Fourth
Order	<i>Zero hours of sleep</i>	<i>Four hours of sleep</i>	<i>Eight hours of sleep</i>	<i>Six hours of sleep</i>
1	<i>sleep</i>	<i>sleep</i>	<i>sleep</i>	<i>sleep</i>
Order	<i>Four hours of sleep</i>	<i>Six hours of sleep</i>	<i>Zero hours of sleep</i>	<i>Eight hours of sleep</i>
2	<i>sleep</i>	<i>sleep</i>	<i>sleep</i>	<i>sleep</i>
Order	<i>Six hours of sleep</i>	<i>Eight hours of sleep</i>	<i>Four hours of sleep</i>	<i>Zero hours of sleep</i>
3	<i>sleep</i>	<i>sleep</i>	<i>sleep</i>	<i>sleep</i>

Ordinal Position of Within-Subjects Conditions

Order	<i>Eight hours of sleep</i>	<i>Zero hours of sleep</i>	<i>Six hours of sleep</i>	<i>Four hours of sleep</i>
4				

Page 160

Time Interval between Treatments

Another way of dealing with order effects is to consider the time interval between the presentation of different conditions. Consider the fatigue effect that arises when participants become exhausted over the course of completing several tasks in succession. Including a rest period may be a simple solution to counteract this fatigue. Similarly, to counteract the contrast effect, a researcher may insert an unrelated task between two conditions. If a condition involves the administration of a drug or some other physiological event that takes time to wear off, the interval between conditions may have to be a day or more. For example, because it takes time to recover from the effects of sleep loss, Roehrs and colleagues (2003) gave participants three to seven days to recover between conditions. A similarly long time interval may be needed with procedures that produce emotional changes such as heightened anxiety or anger. However, introducing an extended time interval may create a separate problem: Participants will have to commit to the experiment for a longer period of time. This can make it more difficult to recruit volunteers, and selective attrition may become a problem as some participants drop out of the experiment. As you can see, designing a carefully controlled experiment free from threats to internal validity is quite a tricky process!

Choosing between Between-Subjects and Within-Subjects Designs

Within-subjects experiments have two major advantages over between-subjects designs: (1) fewer participants are required to complete the experiment, and (2) removing error variance associated with participant characteristics means that any differences between conditions will be easier to detect. These advantages can be very important for some forms of research. However, researchers considering a within-subjects design also need to deal with the challenges of this approach, including order effects. In addition, a within-subjects design may give participants more clues as to the specific hypothesis being studied because they

get to see all possible conditions. Knowing the hypothesis can influence their behaviour, making it difficult to generalize the results to the real world as participants are no longer behaving as they naturally would (see the discussion of *demand characteristics* in [Chapter 9](#)). Between-subjects designs are also preferable when experimental procedures are expected to produce relatively permanent changes, such as surgical procedures like the removal of brain tissue or some psychotherapies.

One way to consider which approach might be most appropriate is to think about how similar these are to what people experience in the real world (Greenwald, 1976). In actual everyday situations, we sometimes encounter variables in a between-subjects fashion: We experience only one condition without a contrasting comparison. For example, if you are interested in jury trials and how the characteristics of a defendant affect jurors, a between-subjects design may be the most appropriate because actual jurors focus on a single defendant in a trial. Real jurors certainly don't evaluate several different defendants with only slight variations in their characteristics. However, for other everyday experiences, we do encounter things in a within-subjects fashion. As an example, when employers consider job applicants, they absolutely do look at the applications of several different applicants one after another. Thus, to study topics like the characteristics of job applicants, a within-subjects design might be more appropriate. Whether to use a between-subjects or within-subjects design may also be partially determined by these issues of external validity ([Chapter 14](#)).Page 161

This chapter explored basic experimental design, including between-subjects and within-subjects designs. In the next chapter, we will consider various issues that arise when you decide how to actually conduct a study.



Illustrative Article: Experimental Design

We are constantly connected. We can be reached by cellphone almost anywhere, at almost any time. Text messages compete for our attention. Email and instant messaging (IM) can interrupt our attention whenever we are using a cellphone or computer. Is this a problem? Most people like to think of themselves as experts at multitasking. But can they be?

A study conducted by Bowman, Levine, Waite, and Gendron (2010) attempted to determine whether IMing during a reading session affected test performance. In this study, participants were randomly assigned to one of three conditions: one where they were asked to IM prior to reading, one in which they were asked to IM during reading, and one in which IMing was not allowed at all. Afterward, all participants completed a brief test on the material presented in the reading.

First, acquire and read the article:

- Bowman, L. L., Levine, L. E., Waite, B. M., & Gendron, M. (2010). Can students really multitask? An experimental study of instant messaging while reading. *Computers & Education*, 54, 927–931.
doi:10.1016/j.compedu.2009.09.024

After reading the article, answer the following questions:

1. This experiment used a posttest-only design. How could the researchers have used a pretest-posttest design? What would the advantages and disadvantages be of using a pretest-posttest design?
2. This experiment used a between-subjects design.
 1. How could they have used a within-subjects design? What would have been the advantages and disadvantages of using a within-subjects design?
 2. How could they have used a matched pairs design? What variables do you think would have been worthwhile to match participants on? What would have been the advantages and disadvantages of using a matched pairs design?
3. What potential confounding variables can you think of?
4. In what way does this study reflect—or not reflect—the reality of studying and test taking in college? That is, how would you evaluate the external validity of this study?
5. How good was the internal validity of this experiment?
6. What were the researchers' key conclusions of this experiment?

7. Would you have predicted the results obtained in this experiment? Why or why not?

Study Terms

Test yourself! Define and generate an example of each of these key terms.

- *between-subjects design* (p. 152).
- *contrast effect* (p. 158).
- *counterbalancing* (p. 158).
- *fatigue effect* (p. 158).
- *Latin square design* (p. 159).
- *matched pairs design* (p. 155).
- *order effect* (p. 157).
- *posttest-only design* (p. 154).
- *practice effect* (p. 157).
- *pretest-posttest design* (p. 154).
- *selection differences* (p. 152).
- *selective attrition* (p. 154).
- *within-subjects design* (p. 152).

Review Questions

Test yourself on this chapter's learning objectives. Can you answer each of these questions?

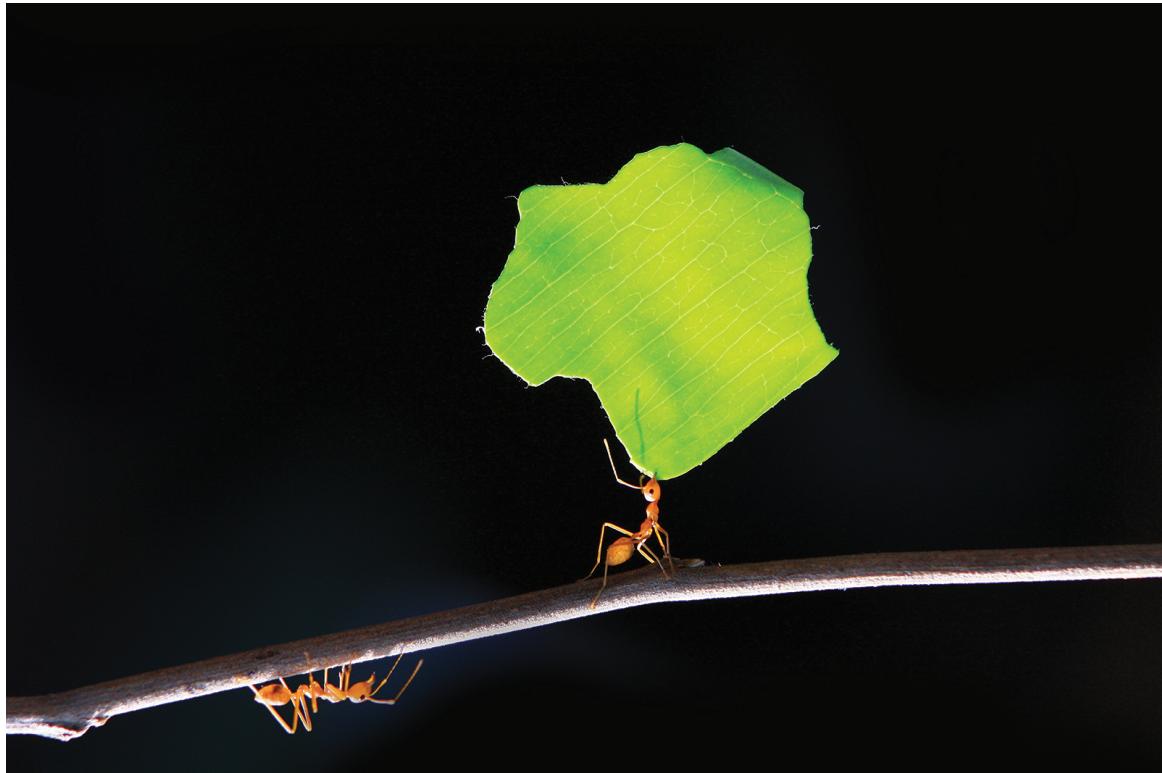
1. What is a confounding variable?
2. What is meant by the internal validity of an experiment? When designing a study, how does a researcher achieve internal validity?
3. How are between-subjects, matched pairs, and within-subjects experiments designed to ensure that groups differ only in terms of the independent variable and nothing more?
4. Describe the differences between a basic between-subjects design and the pretest-posttest design. What are the advantages and disadvantages of each?
5. What is a within-subjects design? What are the advantages and disadvantages of using a within-subjects design versus a between-subjects design?
6. How do researchers overcome the unique challenges of a within-subjects designs?
7. When and why might a researcher decide to use the matched pairs design?
8. What is the difference between random sampling ([Chapter 7](#)) and random assignment?

Deepen Your Understanding

Develop your mastery of these concepts by considering these application questions. Compare your responses with those from other people in your study group.

1. Design a within-subjects experiment that investigates the effect of report presentation style on the grade received for the report. Use two levels of the independent variable: a *professional-style* presentation (high-quality paper, consistent use of margins and fonts, carefully constructed tables and charts) and a *non-professional style* (average-quality paper, frequent changes in the margins and fonts, tables and charts lacking proper labels). Is counterbalancing necessary? Why or why not? Create a diagram illustrating the experimental design. Page 163
2. Professor Pemberton conducted a cola taste test. Each participant in the experiment first tasted two ounces of Coca-Cola, then two ounces of Pepsi-Cola, and finally two ounces of President's Choice brand cola. Participants rated the cola's flavour after each taste. What are the potential problems with this experimental design and the procedures used? Revise the design and procedures to address these problems. Consider several alternatives and think about the advantages and disadvantages of each.
3. Design an experiment to test the hypothesis that single-sex math classes are beneficial to adolescent girls. Operationally define both the independent and dependent variables. Your experiment should have two groups and use the matched pairs design. What variable will you use to match participants? Why? In addition, defend your choice of either a posttest-only design or a pretest-posttest design.

Conducting Studies



©bill2499/123RF.com

Although you now know the basics of the experimental design, there is still a lot of work left to do before you can begin collecting your data and analyze it. This little ant knows a thing or two about hard work, and so do good researchers like yourself.

Learning Objectives

Keep these learning objectives in mind as you read to help you identify the most critical information in this chapter.

By the end of this chapter, you should be able to:

1. [LO1](#) Describe issues to consider when manipulating independent variables.
2. [LO2](#) Describe and give examples of ways to measure dependent variables.
3. [LO3](#) Contrast floor effects and ceiling effects while discussing sensitivity of a dependent variable.
4. [LO4](#) Discuss what it means to “set the stage” for participants.
5. [LO5](#) Describe ways to control participant expectations and experimenter expectations.
6. [LO6](#) Summarize steps for preparing for ethical approval and running a study.

Page 165 Previous chapters have laid the foundation for planning basic studies. Before we consider advanced topics, here we focus on some practical aspects of conducting research. How does a researcher turn a design into an actual study for gathering data? We will follow the process first depicted in [Chapter 1](#), with an emphasis on the middle three steps: (1) finalizing the study design, (2) applying for ethical approval, and (3) collecting data ([Figure 1.1](#)). Although most of these steps apply to experiments as well as non-experiments, some issues are specific to experiments (e.g., manipulating the independent variable). This chapter will be particularly useful if you are conducting a study for a course project, volunteering as a research assistant, completing an honours project, or if you will conduct research in your future career. For everyone, this chapter is intended as a guide to help tie together the practical aspects of the research process presented throughout this book.

Finalizing a Study Design

Once you have a research idea, have stated your hypothesis, and have decided on a research design to test it, the next steps involve detailing the operationalizations for each variable and creating an experience for your participants. For experiments, all efforts must be made to control as many elements of the situation as possible to ensure internal validity ([Chapter 8](#)). All these details of your study should be determined before applying for ethical approval.

To manipulate independent variables in experiments, and to measure variables in any study, you have to construct an operationalization for each variable; that is, you must turn a conceptual variable into a set of operations: specific instructions, events, and stimuli to be presented to participants. In [Chapter 4](#), we considered some examples of operationalizations. Here, we explore various options when creating operationalizations. There are no clear-cut rules for translating conceptual variables into specific operations. Exactly how the variable is operationalized depends on the variable, cost, practicality, ethical concerns, and researcher's creativity.



LO1 Options for Manipulating the Independent Variable in Experiments

In experimental designs, researchers must decide on a way to manipulate the independent variable. One way to think about these operationalizations is to distinguish straightforward manipulations from staged manipulations.

Straightforward Manipulations

Researchers can manipulate a variable simply by presenting material to the participants. These *straightforward manipulations* operationalize independent variables using instructions and stimulus presentations. Stimuli may be presented verbally, in written form, or via video. Let's look at two examples.

Gaucher, Friesen, and Kay (2011) studied the impact of gendered language in job advertisements on job appeal. Both male and female participants were asked to read and rate job ads. In one condition, the ads were worded using stereotypically masculine words (e.g., dominant, independent, determined). In the other condition, those words were replaced with stereotypically feminine ones (e.g., sensitive, collaborate, commitment). To ensure experimental control, all other details about the job were the same. Female participants rated the femininely worded ads as more appealing

than masculinely worded ads, whereas male participants preferred masculinely worded ads. The fact that researchers simply presented participants with the stimulus materials (i.e., the advertisements) makes this a straightforward manipulation. Researchers were able to conclude that job advertisements should be worded very carefully to avoid turning off potentially excellent candidates from particular demographic groups.

Page 166

A great deal of memory research relies on straightforward manipulations. For example, researchers at Dalhousie University showed participants words or pictures on a computer screen (e.g., of a kite) one at a time (Quinlan, Taylor, & Fawcett, 2010). After each word or picture, participants were asked to remember or forget it. These materials (pictures versus words) and specific instructions (remember or forget) constitute two independent variables operationalized using straightforward manipulations. Later, participants completed the dependent variable in the form of a recognition memory task. Overall, pictures were more memorable than words, even when participants had been asked to forget the pictures. These results suggest that intentionally trying to forget an image is more difficult than trying to forget a text-based description of that image. Think of that the next time you are considering whether to read a horror novel or see the movie!

Straightforward manipulations are used to operationalize many independent variables in all areas of behavioural research. Researchers vary the difficulty of material to be learned, the way questions are asked, the characteristics of people to be judged, and various other factors by presenting specific materials to participants and asking them to respond. Whenever the tasks involved in a study mimic experiences and conditions present in everyday life, the study is said to have high *mundane realism*. Reading job advertisements and judging how appealing they are is a task that job seekers do in everyday life (e.g., Gaucher et al., 2011). Deliberately trying to forget information and then trying to remember it is arguably less realistic, however (Quinlan et al., 2010). Therefore, we could argue that the job advertisement study has greater mundane realism than the memory study.

Staged Manipulations

Sometimes it is necessary to create a series of events that occur during the experiment to manipulate the independent variable successfully. These *staged manipulations* can be elaborate situations involving actors. At other times, they simply take the form of a cover story. Deception—with all its ethical concerns we explored in [Chapter 3](#)—is often involved in staged manipulations.

There are two main reasons why staged manipulations are employed. First, the researcher may be trying to create a certain psychological state in participants, such as frustration, anger, or a temporary lowering of self-esteem. For example, researchers have created situations where people believed they were interacting with another person, although the other person's contribution was actually a pre-recorded video (e.g., Cameron, Stinson, Gaetz, & Balchen, 2010). You can see how this operationalization is more complex than a straightforward manipulation, and relies on convincing the participants that they are partaking in a certain kind of situation or interaction.

Second, staged manipulations can be used to simulate situations that occur in the real world. For example, White and Caird (2010) used the University of Calgary Driving Simulator to study the effects of conversing with a passenger on driving errors. Participants were randomly assigned to drive either alone or with a passenger. The passenger, a confederate playing the role of another participant, engaged the real participant in conversation throughout the driving simulation. Relative to the “alone” condition, the presence of this conversation did not affect how much time people spent looking at the road. However, it did reduce participants’ ability to actually notice and respond to others on the road (e.g., pedestrians, motorcycles). The conditions in this staged manipulation simulated common real-world environments that people experience while driving.[Page 167](#)

Staged manipulations sometimes employ a confederate (also known as an accomplice). The nature of this confederate can differ quite widely from study to study. For example, in the driving study above, the confederate was a real person who was present in the situation. In the acceptance study,

in contrast, the confederate's contribution was pre-recorded and presented as a video. Usually, the confederate is introduced as another participant in the experiment. A confederate may be used to create a particular social situation or administer the independent variable. For example, in one study, sedentary women were randomly assigned to exercise on a treadmill alongside a confederate trained to give either positive or negative comments about exercise, depending on which condition the participant was in (Scarapicchia, Sabiston, Andersen, & Garcia Bengoechea, 2013). When the confederate expressed that she enjoyed running, participants chose to work out harder than when the confederate expressed a hatred of running.

The staging involved in the running study was relatively minimal. For a more complex example of staging, consider a study conducted in a bar in the Netherlands (Larsen, Overbeek, Granic, & Engels, 2012). Participants were run in pairs in the campus bar, with the second participant actually being a confederate working for the experimenters. The experimenter offered the pair the chance to get acquainted before starting the study, during which time they could order two drinks paid for by the researchers. The confederate always chose first, and chose drinks that were either non-alcoholic (in one condition) or alcoholic (in the other condition). Participants drank more alcohol when the confederate ordered alcoholic beverages than when the confederate did not, demonstrating the effect of social influence on alcohol consumption, all within a field setting.

Staged manipulations can demand a great deal of ingenuity and some acting ability on the part of confederates. Staged manipulations are used to involve the participants in an ongoing social situation, which the individuals perceive not as an experiment but as a real experience. When a study engages and involves participants in this deep way, the study is said to have *experimental realism*. Note the difference between experimental realism and mundane realism. The tasks in a study might not resemble real-world experiences (i.e., low mundane realism), but they can still engage participants in a meaningful way, producing psychological experiences that are impactful (i.e., high experimental realism). It is often easier to have high experimental realism in staged rather than straightforward manipulations. Researchers assume that the result of an operationalization

with high experimental realism will be natural behaviour that truly reflects the feelings and intentions of the participants. However, staged manipulations often allow for a great deal of subtle interpersonal communication that is hard to put into words. This can make it difficult for other researchers to replicate the experiment. Also, a complex manipulation is difficult to interpret. If many things happened during the experiment, what one thing was responsible for the results? In general, it is easier to interpret results when the manipulation is relatively straightforward. However, the nature of the variable you are studying sometimes demands more complicated procedures, using staged manipulations.

Additional Considerations when Manipulating the Independent Variable

Strength of the Manipulation

When deciding on precise operationalizations for each level of the independent variable, researchers must consider *manipulation strength*. In general, it is a good idea to try to make the manipulation as strong as possible. Essentially, this means making the levels of the independent variable maximally different, while keeping everything else between the two groups the same. This strong manipulation will increase the chances that your results will reveal an effect of the independent variable on the dependent variable, if one really exists. Page 168

To illustrate, consider two different ways University of Manitoba researcher Jessica Cameron and her colleagues (2010) manipulated the riskiness of social interactions in a set of elaborately staged manipulations. One purpose of these experiments was to examine how risk of negative evaluation affects feelings of social acceptance. The risk of negative evaluation for a social interaction was the independent variable, and self-reported feelings of acceptance was the dependent variable. The strength of the manipulation differed across two studies. In one study, the two levels of risk were high risk and no risk, whereas in the other study, the two levels were high risk and less risk. How were these variables operationalized? In the first study, participants were led to believe that another participant

(actually a confederate) had recorded a video to introduce themselves, and the person in this video was very warm and inviting. The video was the same in both conditions. Risk was manipulated by changing how the video was presented to the participant. It was presented as either a response to the participant's own introduction video (i.e., high-risk condition), or as a response to someone else's introduction video (i.e., no-risk condition). In this way, you can see how the first condition entails a much higher risk of rejection since it is a direct response to the person's own very personal introduction. In contrast, the second condition is not a response to the participant's own video, and so there is no real risk of being rejected. In the second study, a much weaker manipulation was examined, comparing a high-risk condition with a less-risk condition (instead of a no-risk condition). Unlike the first study, risk was manipulated by providing people with different information about the confederate. In the high-risk condition, participants were given some basic demographic information about the person. In the less-risk condition, participants were given a statement written by the person that admitted to a serious personal flaw. The difference in manipulations across these two studies illustrate how operationalizations can differ in strength. Across the two studies, the high-risk conditions focused on whether someone likes you, the less-risk condition focused on whether a flawed person likes you, and the no-risk social situation focused on whether someone likes someone else other than you. Not surprisingly, comparing high to low can lead to different conclusions than comparing high to no.

☆ Student Spotlight: Strength of Manipulations ☆

A study conducted by Dr. Fionnuala Murphy when she was an undergraduate helps to illustrate the difference between strong and weak manipulations. In collaboration with her supervisor, Dr. Raymond Klein at Dalhousie University, she conducted a study examining how nicotine affects visual attention. The key manipulation of the independent variable was therefore exposure to nicotine, with the dependent measure involving an attention task. In order to have the strongest manipulation possible, they had some casual smokers complete the attention task directly after smoking a single cigarette. This condition was contrasted with two alternatives: (1) 1 hour after smoking and (2) 24 hours after smoking (with no smoking in the

interim). The contrast in performance between those measured directly after smoking and 24 hours after smoking would involve the strongest manipulation of nicotine exposure (i.e., most versus least). Comparing performance between participants who just finished smoking and those who smoked one hour ago, would be a weaker manipulation. Dr. Murphy is now a scientist working at the University of Cambridge, and the paper was published in the journal *Neuropsychologia* (Murphy & Klein, 1998).

A note of caution: The strongest manipulation possible can sometimes be ethically problematic. Consider the bar study by Larsen and colleagues (2012) described earlier. People might have consumed even more alcohol if the confederate had ten drinks rather than two in the alcohol condition. Yet promoting that much alcohol consumption during an hour may put participants at risk of physical and psychological harm, thereby violating the ethical principle of concern for welfare ([Chapter 3](#)).Page 169

Cost of the Manipulation

Cost is another issue when deciding how to manipulate the independent variable. Some government and private agencies offer grants to support research (e.g., SSHRC, NSERC; [Chapter 3](#)). Because research is often costly, continued public support of these agencies is very important. Researchers who have limited monetary resources may not be able to afford expensive equipment, salaries for confederates, or payments to participants to reduce attrition for longitudinal studies. Also, a manipulation in which participants must be run individually requires more of the researcher's time than a manipulation that allows running many individuals at once. Straightforward manipulations are often much less costly than complex, staged, experimental manipulations.

Manipulation Checks

It is often a good idea to include a [manipulation check](#) to directly measure whether the manipulation of the independent variable was successful, inducing the intended psychological state among participants. Manipulation checks can provide evidence for the validity of your manipulation ([Chapter 5](#)). If you are manipulating anxiety to study its effect on memory, for

example, a manipulation check will tell you whether participants in the high-anxiety group really were more anxious than those in the low-anxiety condition. The manipulation check in this case might involve a self-report of anxiety, a behavioural measure (e.g., number of nervous movements), or a physiological measure (e.g., heart rate change). Because a manipulation check might distract participants or inform them about the purpose of the experiment, it is difficult to decide when to administer it. If the effects of the independent variable are expected to last long enough, manipulation checks can be placed after measuring the dependent variable. However, there is always a risk that the manipulation was successful, but by the time you measure it late in the study, the effects have worn off.

A manipulation check has two advantages. First, if the manipulation check is used in an early pilot study and reveals that your manipulation is not effective, you can change the procedures before running the actual experiment. For instance, if the manipulation check shows that neither the low- nor the high-anxiety group was very anxious, you could change your procedures to increase anxiety further in the high-anxiety condition. In other words, you could increase the strength of your manipulation for the independent variable.

Second, a manipulation check is advantageous if the results show no effect of the independent variable on the dependent variable. In order to conclude that the independent variable truly doesn't affect your dependent variable, you will first need to rule out the possibility that you failed to manipulate the independent variable at all. If both groups are equally anxious after you manipulate anxiety, anxiety can't affect the dependent measure. What if the check shows that the manipulation was successful, but you still do not find an effect? Then you know at least that the results were not due to a problem with the manipulation; the reason for not finding a relationship must lie elsewhere. Perhaps you had a poor dependent measure, or perhaps there really is no relationship between the variables.



LO2 Options for Measuring Variables

We have discussed various issues when it comes to measuring variables, including reliability, validity, and reactivity. Here, we introduce some additional practical concerns when operationalizing variables that are measured, starting by outlining three broad categories of measures: self-report, behavioural, and physiological. [Table 9.1](#) provides a summary. Page 170

Table 9.1 Some Ways to Operationalize Measured Variables

Type of Measure	Examples of Techniques or Tools	Examples of Variables Operationalized Using These Techniques/Tools
Self-report	Paper-and-pencil questionnaire	Attitudes about something
	Face-to-face interview	Intentions to do something
	Online questionnaire	A person's values, self-esteem, mood, anxiety, relationship satisfaction, personality traits

Type of Measure	Examples of Techniques or Tools	Examples of Variables Operationalized Using These Techniques/Tools
Behavioural	Audio or video recorder	Self-control (amount of ice cream eaten, length of persistence on boring task) Creativity (number of ideas generated per minute)
	Eye tracker	Reaction time (speed of detecting a flashing light) Facial expression of emotion (coded photographs)
	Electronic activated recorder	Attention (eye tracker, number of hazards avoided in driving simulator)
	Weight scale	Liking (distance seated apart from someone)
	Still camera	Efficacy of a bulimia intervention (weight gained or lost) Memory (number of items recalled) Generosity (amount of money donated)
	GSR, EMG, ECG, EEG	
Physiological	Blood analysis	Stress (sweating from GSR, cortisol in saliva) Genetic marker for mental illness (blood analysis)
	Saliva analysis	Physical fitness (heart rate change during exercise)
	Heart rate	Size of amygdala or damage to hippocampus (MRI)
	Breathing rate	Brain activation when looking at image of romantic partner (fMRI)
	Blood pressure	
	MRI, fMRI	

Self-Report Measures

A self-report measure can be used to measure many different aspects of human thought and behaviour, including explicit attitudes, judgments about someone's personality characteristics, intended behaviours, emotional states, and confidence in one's judgments. Responses are typically made

using rating scales with descriptive anchors. For example, in the self-esteem and social acceptance study described earlier in this chapter, participants were asked five questions about perceived social acceptance (e.g., “The other participant probably likes me.”) All questions used a 7-point scale, like this one:

1	2	3	4	5	6	7
			Neither			
Strongly Disagree	Disagree	Slightly Disagree	Agree nor disagree	Slightly Agree	Agree	Strongly Agree
			Disagree			

When using self-report measures, using a published scale that has been validated is typically preferred ([Chapter 5](#)), or you can create your own measure ([Chapter 7](#)).

Behavioural Measures

A [*behavioural measure*](#) is a direct observation of behaviours. As with self-reports, there is an almost endless number of possible behaviours you could choose to observe and record. There are also many different aspects of a behaviour that one can record. For example, a researcher may choose to note whether a given behaviour occurs (e.g., whether someone helps, makes an error on a test, chooses one activity over another). You can also record the number of times a behaviour occurs, often within a given time period: This is known as the *rate* of a behaviour. Another option is to measure how quickly a response occurs after a stimulus, known as a *reaction time*. There are several other aspects of behaviour that you can also measure, such as how long a behaviour lasts (i.e., duration). Deciding which aspect of behaviour to measure depends on which is most theoretically relevant, or which aspect of behaviour is most likely to be affected by the independent variable. [Chapter 6](#) provides a deeper exploration of how to code behaviours and count them.

Physiological Measures

A *physiological measure* is a recording of a response of the body. Many such responses are available (see Cacioppo, Tassinary, & Berntson, 2007), and some examples should help illustrate this idea. One popular form of physiological measure is the *galvanic skin response* (GSR): the electrical conductance of the skin, which changes when sweating occurs. This is often used to measure general emotional arousal, anxiety, or stress. An *electromyogram* (EMG) measures muscle tension, and the *electrocardiogram* (ECG) measures heartbeat regularity and rate. Both can be used as indicators of tension or stress. EMG can also be used to measure which muscles in the face are being activated, as an indicator of different emotional expressions. The *electroencephalogram* (EEG) is a measure of the electrical activity of brain cells. It can be used to measure activity in different parts of the brain as learning occurs, or brain activity during different stages of sleep. Measurements of the brain can also be made using *magnetic resonance imaging* (MRI), which captures images of brain structures. This allows scientists to compare brain structures between people with a particular condition (e.g., schizophrenia) and those without the condition. A related technology, *functional MRI* (fMRI), measures blood flow in the brain to make inferences about which regions are involved while a participant performs a specific task. Physiological measurements can also be taken from various bodily fluids. For example, the stress hormone cortisol is measured in samples of saliva (e.g., Pruessner et al., 2013). Blood can also be sampled and analyzed, to reveal genetic aberrations associated with mental illness, for example (e.g., Uher & Weaver, 2014).

☆ Student Spotlight: Using MRI for Measurement ☆

Have you ever had an MRI scan? For many people, lying on that bed and being shuttled into the big metal donut can be a stressful experience. Hanah Chapman's undergraduate research project directly examined how anxiety levels change for a person's first scan versus that person's second scan, as well as how anxiety changes over the course of a single scan. This is an important topic because anxiety affects brain activity and could pose problems for interpreting the results from neuroimaging studies. Working with Dr. Benjamin Rusak and Dr. Denise Bernier at Dalhousie University, they measured anxiety levels during two different MRI scans as well as

during a control period outside of the scanner. You can read more about the results of their fascinating study in the journal *Psychiatry Research* (Chapman, Bernier, & Rusak, 2010).

Sometimes the nature of the variable being studied requires a very particular type of measure, but in other cases, the same variable can be measured using more than one category of measurement. A measure of helping is almost by definition a behavioural measure, whereas personality traits are almost always measured using self-report. For many variables, however, all three types of measurement might be possible and appropriate (i.e., self-reports, behavioural measures, and physiological measures). Interpersonal attraction could be measured with a self-report rating scale or with a behavioural measure such as the amount of time spent looking into one another's eyes. Just as using a multitude of research approaches will lead to the strongest conclusions (i.e., a multi-method approach), employing a diversity of measurement types is also a good idea. Note that these different measurement types need not appear in the same study; they can also be used across a series of studies.

Page 172



LO3 Additional Considerations when Measuring Variables

Sensitivity

Recall that it is typically important to create strong manipulations of independent variables, maximizing the differences between groups. Likewise, the dependent variable should be sensitive enough to detect any resulting differences between groups. A self-report measure of liking that asks, “Do you like this person?” with a simple “yes” or “no” response format is less sensitive than one that asks, “How much do you like this person?” on a 5- or 7-point scale. With the first measure, people may say yes even if they have some negative feelings about the person. The second measure is more sensitive by enabling a gradation of liking and such a scale would make it easier to detect differences in the amount of liking. Consider another example: reaction time. Changes in reaction time typically occur at the level of milliseconds (i.e., thousandths of seconds). A timer that measures only to the nearest second would not be sensitive enough to detect small differences in reaction time across conditions.

The issue of *sensitivity* is particularly important when using behavioural measures of performance. There are many different ways to measure cognitive performance. Memory, for example, can be measured using recall, recognition, or reaction time. Such tasks vary in their difficulty. Sometimes a task is so easy that everyone does well. This results in what is called a *ceiling effect*: if everyone does well then there isn’t much variability in the scores and so the measure lacks sensitivity to detect differences. This can lead to problems interpreting the results. An independent variable might appear to have no effect on the dependent measure, but this is actually because the performance of all participants is at ceiling. The opposite problem occurs when a task is so difficult that hardly anyone can perform well: This is called a *floor effect*. Ideally, we want the scores for our measures to be not too high or too low and to have lots of variability (i.e., a good spread of scores), exhibiting the sensitivity to detect differences. To diagnose a potential ceiling or floor effect, look for average values close to the minimum or maximum possible score.

Multiple Measures

Because any variable can be measured using various operationalizations, researchers sometimes include multiple measures of the same variable. In an experiment investigating the stress of music performances among young

children, for example, researchers included three measures of stress: (1) children's self-reported stress about an upcoming performance, (2) cortisol levels based on saliva samples taken immediately after the performance, and (3) observer ratings of anxious behaviours during the performances (Boucher & Ryan, 2011). If the independent variable has the same effect on several measures of the same dependent variable, our confidence in the results is increased. It is also useful to know whether the same independent variable affects some measures of a dependent variable but not others. Researchers can also be interested in studying effects for several entirely different dependent variables. For example, an experiment on new active-learning techniques might examine academic performance, quality of interaction among classmates, and teacher satisfaction.

Page 173

When you have more than one measure, the question of what order to sequence them arises. Is it possible that responses to one measure will be different if it comes before or after another measure? This issue is similar to the order effects we discussed for within-subjects designs ([Chapter 8](#)). We are worried about the same issues as before, carry-over effects, as well as any influence of fatigue. There are two ways to deal with this issue. One strategy is to present the most important measures first and the less important ones later. This way, order will not be a problem in interpreting the results for the most important variables. Even though order may be a potential problem for some of the measures, the overall impact on the study is minimized. Another strategy is to counterbalance the order of presenting the measures ([Chapter 8](#)), or rely on complete randomization of order.

Making multiple measurements in a single study can be efficient. However, in some cases it may be necessary to conduct a series of studies to explore the relationships between one variable and many others. Most importantly, to ensure ethical, open disclosure whenever multiple measures are used, it is important to include all of them in the research report, even if results are mixed ([Chapter 3](#)). It is absolutely unethical to report only what happened with the measures that "worked," or to only report the results that confirm your hypothesis.

Cost of Measures

Some measures are more expensive than others. Self-report questionnaire measures are generally inexpensive or even free, whereas measures that require trained observers or elaborate equipment can become quite costly. A researcher studying attention, for example, might use an eye-tracking device to record each participant's eye movements while watching a film clip. This would entail purchasing and learning how to use specialized equipment and software. This means expenses in terms of both equipment and personnel. Often, researchers need resources from the university or outside agencies to carry out such research.



LO4 Setting the Stage

Once you have decided on specific operationalizations for all variables, consider asking another researcher for feedback on your design before moving forward. Research is most successful when it is a collaborative enterprise, and it is always useful to get other perspectives on your design decisions. The next step will then be to plan the experience from the participant's viewpoint, or "set the stage" (Aronson, Brewer, & Carlsmith, 1985). There are no clear-cut rules for setting the stage, except that the study's setting must seem plausible to the participants. What is the exact procedure? How will you present and explain the sequence of tasks to participants? Will the participants fully understand what you are asking them to do?

In most cases, you will need to prepare the informed consent form ([Chapter 3](#)) and explain to participants why the experiment is being conducted. Sometimes the rationale given is completely truthful, although only rarely will you want to tell participants the actual hypothesis (see demand characteristics, [below](#)). For example, you might say in a general way that you are conducting an experiment on memory, when you are actually studying a specific aspect of memory such as working memory capacity. If you decide that any deception is necessary, you will need to plan a debriefing session at the end of the experiment (see [below](#)).

If collecting data online, prepare the website, including a welcoming message and a closing message thanking participants for their time. If collecting data in person or over the telephone, we recommend preparing a step-by-step script that starts with welcoming the participant to the study and finishes with debriefing and thanking the participant. Ensure that anyone who will be acting as an experimenter has practised using the script and is comfortable using it. Practice is important to ensure experimental control, especially when using elaborately staged manipulations. Once you have prepared all the materials and set the stage for participants, there are just a few more issues to consider before applying for ethical approval from your institution's Research Ethics Board.



LO5 Advanced Considerations for Ensuring Control

Good research design means eliminating as many alternative explanations for the results as possible. For example, researchers want to avoid confounding variables: variables not of interest that cannot be separated from the independent variable (i.e., covary with levels of the independent variable) and might explain the results ([Chapter 4](#)). Additional control procedures may be necessary to address other types of alternative explanations. Two of these issues concern expectations on the part of both the participants and the experimenters.

Controlling for Participant Expectations

Demand Characteristics

Sometimes experimenters do not wish to inform participants about the specific hypotheses being studied or the exact purpose of the research. The

reason for this lies in the problem of *demand characteristics* (Orne, 1962). A demand characteristic is any feature of a study that might inform participants of the study's purpose and consequently affect their behaviour. When participants form particular expectations about the hypothesis of the study, they might deliberately act in ways to confirm the hypothesis or even undermine this hypothesis. Although participants tend to act cooperatively, not everyone does. Nichols and Maner (2008) found that participants who had been told the hypothesis tended to act in ways that confirmed it, especially among those who liked the experimenter.

One way to control for demand characteristics is to use deception to mislead participants about the purpose of the study. An experimenter can devise elaborate cover stories to explain the purpose of a study and to disguise what is really being studied (e.g., Laney et al., 2008). The researcher may also attempt to disguise the dependent measure by using an unobtrusive measure, or by placing the measure among a set of unrelated *filler items* in a questionnaire. Another approach is to assess whether demand characteristics are a problem by asking participants what they thought the research was about (e.g., Laney et al., 2008). It may be that participants do not have an accurate view of the purpose of the study and so are unlikely to be reacting to demand characteristics. If some individuals do guess the hypotheses of the study, their data can be analyzed separately to see if this knowledge influenced their responding.



Think about It!

How are demand characteristics related to the issue of reactivity, discussed in the context of measurement in [Chapter 5](#) and for observational approaches in [Chapter 6](#)? How can you apply what you learned about reducing reactivity in those contexts to reduce demand characteristics?

Demand characteristics may be eliminated when people are not aware that an experiment is taking place or that their behaviour is being observed. Thus, observational research in which the observer is concealed or is using

unobtrusive measures can minimize the problem of demand characteristics. Page 175

Placebo Effects

A special kind of participant expectation arises in research on the effects of treatments, including the effects of drugs. Consider an experiment investigating whether an antidepressant reduces depression. People who have been diagnosed with depression are randomly assigned to receive the drug or nothing at all. Now, suppose that the drug group shows an improvement. We do not know whether the improvement was caused by the properties of the drug or by what participants expect to feel after taking the drug. This is known as a *placebo effect*. In other words, just administering a pill, or other form of treatment, may be sufficient to cause an improvement in behaviour. To control for this possibility, a *placebo group* can be added. To continue the drug example, participants in the placebo group receive a pill or injection containing an inert, harmless substance (e.g., a sugar pill); they do not receive the drug given to the experimental group. Those in the placebo control group may still show an improvement (i.e., a placebo effect), but if the active properties of the drug actually work, then the experimental group should show greater improvement than the placebo group. If the placebo group improves as much as the experimental group, then the improvement observed due to the experimental drug is likely just a placebo effect. Placebo control groups are not limited to drug interventions, however. A placebo effect can occur during any experimental condition in which participants have expectations of some effect or change. For example, a clinical intervention involving talk therapy would also require a placebo control group, people who are given just as strong expectations of improvement.

Sometimes, participants' expectations are the primary focus of an investigation. For example, Darredeau and Barrett (2010) conducted an experiment to determine whether nicotine inhalers reduce cigarette cravings because they contain nicotine or because users expect them to reduce cravings. The experimental design had four groups: (1) given nicotine—told nicotine, (2) given no nicotine—told no nicotine, (3) given nicotine—told no nicotine, (4) given no nicotine—told nicotine. This design is called

a *balanced placebo design*. People who believed they had inhaled nicotine (Groups 1 and 4) reported very similar intentions to reduce smoking, although people in Group 4 were not actually given any nicotine. Believing nicotine was inhaled was more important than actually inhaling nicotine for reducing future intentions to smoke.

In some areas of research, the use of placebo control groups has ethical implications. Suppose you are studying a treatment that really has a positive effect on people (e.g., reducing symptoms of depression). In this case, it is important to help those people who are in the control conditions. For example, control participants may be given the treatment after the study is completed, known as a “waitlist control condition.” This is because these control participants are effectively on a “waitlist” for the real treatment.

Placebo effects are a serious concern for many areas of research. There has been much research and debate on the extent to which the beneficial effects of antidepressants are due to placebo effects. Two meta-analyses (i.e., studies that combine the results of many single experiments; [Chapter 14](#)) indicate that antidepressants may not have much greater effect than a placebo among people with mild or moderate levels of depression, but the medication works better than a placebo among those with severe depression (Fournier et al., 2010; Kirsch et al., 2008).

Controlling for Experimenter Expectations

Experimenters are usually aware of the purpose of the study and likely have expectations about how participants should respond. These expectations can bias the results, a problem known as [*experimenter bias*](#), or *experimenter expectancy effects* (Rosenthal, 1967, 2003).Page 176

Experimenter bias may occur whenever the experimenter knows which condition the participants are in. There are two potential sources of experimenter bias. First, the experimenter might unintentionally treat participants differently depending on what condition they are in. Certain words might be emphasized when reading instructions to one group but not the other, or the experimenter might smile more when interacting with people in one of the conditions. These differences could alter the behaviour

of participants in one condition more than the other, creating a difference in conditions. The second source of bias can occur when experimenters record participants' behaviours, with subtle differences emerging in how the experimenter interprets and records behaviours for people in different conditions. Once again, this can result in differences in scores emerging between conditions that are a function of experimenter bias rather than the independent variable.

Research on Expectancy Effects

Hundreds of studies have been conducted investigating the impact of expectations by experimenters, teachers, interviewers, and so on (Rosenthal, 2003). Perhaps the earliest demonstration of the problem is the case of Clever Hans, a horse whose alleged brilliance was revealed by Pfungst (1911) to be an illusion. Rosenthal (1967) describes Clever Hans:

- Hans, it will be remembered, was the clever horse who could solve problems of mathematics and musical harmony with equal skill and grace, simply by tapping out the answers with his hoof. A committee of eminent experts testified that Hans, whose owner made no profit from his horse's talents, was receiving no cues from his questioners. Of course, Pfungst later showed that this was not so, that tiny head and eye movements were Hans' signals to begin and to end his tapping. When Hans was asked a question, the questioner looked at Hans' hoof, quite naturally so, for that was the way for him to determine whether Hans' answer was correct. Then, it was discovered that when Hans approached the correct number of taps, the questioner would inadvertently move his head or eyes upward—just enough that Hans could discriminate the cue, but not enough that even trained animal observers or psychologists could see it. (p. 363)

An example of more systematic research on expectancy effects is a classic study by Rosenthal (1966). In this experiment, graduate students trained rats that were described as coming from either "bright" or "dull" genetic strains. The animals actually came from the same strain and had been randomly assigned to the bright and dull categories. However, the "bright" rats did end up performing better than the "dull" rats. Subtle differences in

the ways the students treated the rats or recorded their behaviour must have caused this result.

If horses and rats can respond to subtle cues, it is reasonable to suppose that humans can too. Experimenter bias can be communicated to humans by both verbal and non-verbal means (Doyen, Klein, Pichon, & Cleeremans, 2012; Jones & Cooper, 1971). In one study, experimenters were told during their training that a manipulation would either cause participants to walk more quickly or more slowly (Doyen et al., 2012). It seems that experimenters unintentionally influenced participants' responses: Participants walked more slowly when experimenters expected them to do so than when they did not.

Expectations can also influence evaluations of behaviour, as illustrated in one experiment that used a straightforward manipulation (Bruchmüller, Margraf, & Schneider, 2012). Therapists were mailed a realistic description of a case study, in which an adolescent had enough symptoms to suggest attention-deficit/hyperactivity disorder (ADHD), but did not fully meet diagnostic criteria. This adolescent was more likely to be misdiagnosed with ADHD when presented as a boy than as a girl, even though all other information was the same. This means that the therapists' expectations regarding gender affected their diagnoses. Page 177

Solutions to the Expectancy Problem

There are a number of ways to address experimenter bias. One solution is to run everyone in all conditions simultaneously, so that the experimenter's behaviour is exactly the same for all participants. This solution is feasible only under certain circumstances, such as when the study relies on printed materials and the experimenter's instructions to participants are the same for everyone. Alternatively, researchers can use computer survey software to administer the independent variables and record responses. Such automated procedures leave little room for experimenter expectations to influence results.

Other solutions target the experimenters. All experimenters should be well-trained and should practise behaving consistently with all participants.

When they are particularly concerned about expectancy effects, researchers will use experimenters who are unaware of the hypothesis. This means the person conducting the study or making observations is blind to what is being studied or which condition the participant is in. In a [single-blind procedure](#), the participants are unaware of which condition they are in (e.g., whether a placebo or the actual drug is being administered). In a [double-blind procedure](#), neither the participant nor the experimenter knows the participant's condition. Double-blind procedures usually require two different experimenters: one who administers the independent variable, and another who takes over and administers the dependent variable, without knowing what independent variable was assigned.

Because of the problem of experimenter bias in the form of expectancy effects, solutions such as the ones described should be incorporated into experimental procedures. In addition, there are ways to design studies to specifically measure the expectancy effects of experimenter bias (see Klein et al., 2012). Sometimes these effects can be difficult to anticipate. Ask experienced colleagues to read your study's procedures or experience them as a mock participant. Their feedback may help you avoid expectancy effects, demand characteristics, or other problems before collecting data. Once your study is fully designed and materials are prepared, the next step before collecting data is to seek ethics approval.



LO6 Seeking Ethics Approval

Ethical concerns must be kept in mind throughout the research process. Recall from [Chapter 3](#) that before collecting data, researchers must seek approval from their institution's Research Ethics Board (REB). The REB will examine the procedure, materials, and informed consent form. Deception and other anticipated risks must be explained and justified, and weighed against potential benefits. Researchers must also explain how they will ensure confidentiality and anonymity (if possible). In this section, we will emphasize two additional decisions that usually must be made before applying for ethics approval: (1) the participant selection process and (2) debriefing procedures. [Table 9.2](#) offers a checklist summarizing common decisions that need to be made before applying for ethics approval. Consult your institution's REB to ensure you meet their criteria.

Table 9.2 Checklist: What to Do before Applying for Ethics Approval

Operationalized your independent variable (if applicable)

- _____ *Operationalized all measures (including dependent variables)*
- _____ *Sought feedback from colleagues on the method (e.g., to avoid confounds)*
- _____ *Listed the exact procedure each participant will experience*
- _____ *Created all materials participants will use (e.g., online or paper-and-pencil questionnaire)*
- _____ *Created the informed consent form*
- _____ *Planned a way to debrief participants (if necessary)*
- _____ *Justified who will be included and excluded from participating (if targeting a particular population, explain why)*
- _____ *Determined the number of participants to be run in the study*
- _____ *Prepared participant recruitment materials (e.g., poster, e-mail)*
- _____ *Determined how confidentiality and anonymity will be maintained during and after data collection*
- _____ *Minimized any foreseeable risks; noted any foreseeable benefits*
- _____ *Completed institutional Research Ethics Board's application form*

Selecting Research Participants

The population of interest for your study may be children, university students, elderly adults, rats, pigeons, primates, or even bees or flatworms. In all cases, the participants or subjects must somehow be selected. The method used to select participants must be justified to the REB (in the case of humans) or the Animal Care Committee (in the case of non-human animals; [Chapter 3](#)), and has implications for generalizing the research results from your sample to a population. In the case of human participants, the procedures you plan to use to recruit participants will likely need to be approved by your REB (to ensure they are not coercive or misleading).

Prepare these procedures thoughtfully before applying for approval. Page 178

Recall from [Chapter 7](#) that most research projects involve sampling research participants from a population of interest. A researcher might be interested in the population of people who suffer from schizophrenia, or people who are bilingual, or all Canadians living in Canada. Samples may be drawn from any population using probability sampling (e.g., random sampling) or non-probability sampling ([Chapter 7](#)). Sampling is important as it informs whether we can generalize the results to a population, and if so, which populations ([Chapter 14](#)).

Whenever a specific population is targeted, it must be justified. Recall from our discussion of ethics that the principle of justice requires that the benefits and burdens of research are fairly distributed ([Chapter 3](#)). Special procedures are needed when studying members of sensitive populations, such as Indigenous peoples, people with mental illnesses, children, or people living in poverty or in institutions (e.g., prisons). Consult your institution's REB for advice on conducting such research ethically.

How many participants will you need in your study? In general, increasing your sample size increases the likelihood that you will find an effect, assuming there is one to be found. Larger samples provide more accurate estimates of population values ([Table 7.2](#)), and so are more likely to be representative of the populations from which they are drawn. A formal approach to selecting a sample size is discussed in [Chapter 13](#).

Planning the Debriefing

After all data are collected, a debriefing session provides an opportunity for the researcher to explain the ethical and educational implications of the study, verbally and/or in writing. Written debriefing forms are common in research conducted online. Debriefing sessions should always occur whenever any form of deception has been used ([Chapter 3](#)). These sessions include an explanation of why the deception was considered necessary, reassurance that believing the deception does not reflect poorly on the person, and an apology to attempt to repair any negative feelings.

Debriefing sessions are also especially important whenever participants are put at risk in any way. If the study is expected to trigger psychological disturbance, such as negative feelings or troubling thoughts, researchers may offer contact information for campus counselling services. The debriefing is an important opportunity for participants to raise any concerns they may have about the study, and for the researcher to repair any negative moods that may have arisen as a result of participation. It is very important that participants leave the study feeling as good or better than when they arrived.

Page 179

The debriefing can also provide an opportunity to learn more about what participants were thinking during the study, their thoughts on the purpose of the study. After experiments, participants can be asked how they interpreted the independent variable manipulation, and what they were thinking when they responded to the dependent measures. Chartrand and Bargh (2000) offer a specific set of questions to probe for suspicion after using deception, called a *funnelled debriefing*. These questions begin broadly, but then narrow in on the key deception, just like a funnel. The resulting information can prove useful in diagnosing whether participants were behaving as they normally would, or if they were suspicious of the manipulated independent variable.

During a debriefing, researchers may also ask participants to refrain from discussing the study with others. Such requests are typically made when more people will be participating. People who have already participated are aware of the general purposes and procedures, and so it is often important that these individuals avoid revealing this information to potential future participants.

Collecting Data

Once you have received ethical approval, you are almost ready to collect data! Examine [Table 9.3](#) for a final checklist to be sure you are ready for your first participants. There are two additional issues to keep in mind: (1) whether to complete a pilot study and (2) the commitments that researchers have to the participants.[Page 180](#)

Table 9.3 Checklist: What to Do before Collecting Data

- Determined who will collect data (e.g., yourself, research assistants)*
- Created a script that all experimenters will follow when interacting with each participant*
- Received feedback from another researcher about the method, script, and materials (e.g., to avoid demand characteristics)*
- Practised the script with all experimenters until everyone is comfortable*
- Avoided telling the experimenters the hypothesis (if concerned about experimenter bias)*
- Ensured that experimenters know and follow procedures for maintaining participant confidentiality and anonymity*
- Determined exactly how participants will be randomly assigned to condition, including any single- or double-blind procedures*
- Prepared all necessary materials (e.g., photocopies, online surveys)*
- Determined/booked the location and times for data collection*
- Received ethical approval from your institution's Research Ethics Board*

_____ *You should now be ready to collect data!*

Pilot Studies

When procedures are particularly elaborate or costly, or when there will be only a single opportunity to collect data, researchers sometimes choose to conduct a *pilot study*. A pilot study is a “trial run,” with a small number of participants, to test out the procedures. This sample is typically drawn from the same population as the sample the researcher ultimately hopes to test. Because data are collected from participants, the pilot study must be included in the ethics application.

A pilot study will often reveal whether participants understand the instructions, whether the experimental setting seems plausible, whether any questions are confusing, and so on. This process can be especially important when using a staged manipulation of the independent variable, to ensure that the scenario is meaningful and believable. Sometimes, pilot participants are questioned in detail about their experience during the experiment. Another method is to use a “think aloud” protocol (described in [Chapter 7](#)), in which pilot participants verbalize their thoughts about what is happening during the study. Such procedures provide researchers with information about participant experiences, allowing them to make any necessary changes in the method before conducting the real study. Moreover, a pilot study allows experimenters who are collecting the data to become comfortable with their roles and to standardize their procedures. If you wish to make any major changes to the method or materials after the pilot study, submit an amendment to your original ethics application for updated approval.

Researcher Commitments

Researchers make several implicit “contracts” with participants during the course of a study. For example, if participants agree to be present for a study at a specific time, it is crucial that the researcher or research assistant collecting data is there. Participants value punctuality, listing it as an important obligation of researchers (Epstein, Suedfeld, & Silverstein,

1973). Similarly, if researchers promise to send a summary of results to participants, they should do so. If participants are to receive course credit for participation, the researcher must immediately follow through on this promise. These are little details, but they are very important in maintaining trust between participants and researchers, thereby supporting the advancement of science.

What Comes Next?

Analyzing and Interpreting Results

After the data have been collected, the next step is to analyze them. Statistical analyses allow the researcher to examine and interpret the pattern of data obtained in the study. Are the variables related to one another? Does the independent variable have an effect on the dependent variable? We will explore basic statistical concepts in [Chapters 12](#) and [13](#), with some additional calculations provided in [Appendix B](#). Depending on the results of the study, the researcher might choose to conduct a follow-up study to see if the results can be replicated using a new sample of participants ([Chapter 14](#)), to rule out alternative explanations, or to deal with problems in the first study.

Communicating Research to Others

The final step is to write a report that details why you conducted the research, how you obtained the participants, what procedures you used, and what you found (refer to [Chapter 2](#) for the common APA style format and [Appendix A](#) for further tips). Researchers report their results at scientific conferences, and submit them for publication in journals.Page 181

Professional Conferences

National and regional professional associations such as the Canadian Psychological Association (CPA), the American Psychological Association (APA), and the Association for Psychological Science (APS) hold annual meetings at which psychologists and psychology students present their research and learn about research being done by their colleagues. Journalists also attend, so that they can write articles communicating the latest research to the public. Frequently, researchers deliver oral presentations to an audience. Poster sessions are also common, in which

researchers display posters summarizing research findings and are available to discuss the results. See [Appendix A](#) for an example poster layout.

Journal Articles

There are many journals in which research papers are published ([Chapter 2](#)). When a researcher submits a paper to a peer-reviewed journal, the editor sends it for peer review: Two or more scientists read the paper and provide an evaluation. These evaluators can either recommend acceptance (often with the stipulation that revisions be made) or rejection (which is much more common than acceptance). As many as 90 percent of papers submitted to the more prestigious journals are rejected. Many of these rejected papers are submitted to other journals and eventually accepted for publication, but much research is never published.

This chapter covered the process that researchers take to convert their research designs into studies ready for participants to experience. As you can see, there are many decisions that must be made when operationalizing variables and developing procedures, applying for ethical approval, and preparing experimenters to interact with participants. No study will ever be designed or executed perfectly, but careful planning, along with advice from experienced colleagues, will help you develop a study that allows for meaningful inferences.



Illustrative Article: Conducting Experiments

When a person feels safe, they counter-intuitively take more risks. This has been observed among people driving cars, kids on a playground, and bicyclists. This is a problem: The point of safety equipment is not to encourage people to behave less safely! The safety equipment used in past research was designed to be used with the activity that participants completed for the study (e.g., bike helmet with riding a bike). Would the increased risk-taking occur even if the equipment was not designed for the activity?

Gamble and Walker (2016) attempted to address this question with an experimental manipulation and some clever deception by attempting to credibly convince research subjects that they were taking part in an eye-tracking experiment that required either a baseball cap or a helmet. They found that participants who wore the helmet showed higher levels of both risk-taking and sensation-seeking compared to those who wore the cap.

First, acquire and read the article:

- Gamble, T., & Walker, I. (2016). Wearing a bicycle helmet can increase risk taking and sensation seeking in adults. *Psychological Science*, 27, 289–294. doi:10.1177/0956797615620784

Page 182 Then, after reading the article, consider the following:

1. What is the basic design of this study?
2. Describe the manipulation of the independent variable, and identify the manipulation as straightforward or staged.
3. In this chapter, we discuss three types of dependent measures: self-report, behavioural, and physiological. In the experiments presented in this paper, what types of dependent measures were used? Could other types of dependent measures have been used? Explain.
4. Do you think that the risk of overinflating a balloon equates to the risks associated with riding a bicycle? Explain why it does or why it does not.
5. These researchers did not use any manipulation checks in their experiments. How would you design a manipulation check for this experiment?
6. Is the deception used in this study ethical?
7. Evaluate the internal validity of this study.
8. Do you believe that their conclusion is justified, given the experiment and results?

Study Terms

Test yourself! Define and generate an example of each of these key terms.

- behavioural measure (p. 170).
- ceiling effect (p. 172).
- demand characteristics (p. 174).
- double-blind procedure (p. 177).
- experimental realism (p. 167).
- experimenter bias (p. 175).
- filler items (p. 174).
- floor effect (p. 172).
- manipulation check (p. 169).
- manipulation strength (p. 167).
- mundane realism (p. 166).
- physiological measure (p. 171).
- pilot study (p. 180).
- placebo group (p. 175).
- self-report measure (p. 170).
- sensitivity (p. 172).
- single-blind procedure (p. 177).

- *staged manipulations* (p. 166)
- *straightforward manipulations* (p. 165).

Review Questions

Test yourself on this chapter's learning objectives. Can you answer each of these questions?

1. Compare and contrast staged versus straightforward options for manipulating an independent variable.
2. Why is it important to consider the strength of the manipulation for an independent variable? How might a researcher determine whether it is strong enough?Page 183
3. Contrast three different ways to measure dependent variables.
4. What does it mean to say that a dependent variable is sensitive? What are ceiling and floor effects?
5. What are demand characteristics? Describe ways to minimize demand characteristics.
6. What does “setting the stage” involve? How might it determine the content of the debriefing?
7. What are experimenter bias effects (i.e., experimenter expectancy)? What are some solutions to the experimenter bias problem?
8. What are all the decisions that need to be made before applying for ethical approval? Are pilot studies included in the ethics application?
9. How would you train someone to be the experimenter for your study?

Deepen Your Understanding

Develop your mastery of these concepts by considering these application questions. Compare your responses with those from other people in your study group.

1. Dr. Turk studied the relationship between age and reading comprehension, predicting that older people will show weaker comprehension than younger people. Groups of participants who were 20, 30, 40, and 50 years old read a chapter from a book written by physicist Stephen W. Hawking (1988), entitled *A Brief History of Time: From the Big Bang to Black Holes*. After reading the chapter, participants were given a comprehension measure. No relationship was observed between age and comprehension scores, and all age groups had equally low comprehension scores. Why do you think no relationship was found? Identify at least two possible reasons.
2. Revisit the experiment on facilitated communication by children with autism that was described in [Chapter 2](#), as an example of using past research to generate ideas (Montee, Miltenberger, & Wittrock, 1995). Interpret the findings of that study in terms of experimenter expectancy effects.
3. Your lab group has been assigned the task of designing an experiment to investigate the effect of time spent studying on a recall task. Thus far, your group has come up with the following plan: “Participants will be randomly assigned to two groups. People in one group will study a list of five words for five minutes, and those in the other group will study the same list for seven minutes. Immediately after studying, participants will read a list of ten words and circle those that appeared on the original study list.” Make at least two improvements to this experiment, and explain why those changes are useful.
4. Design an experiment using a staged manipulation to test the hypothesis that when people are in a good mood, they are more likely

to contribute to charity. Include a manipulation check in your design.

Chapter 10

Page 184

Research Designs for Special Circumstances



©Corey Hochachka/Design Pics

Unique contexts and situations, like a tiny frog phoning you on a landline, call for some unique research tools. Let's learn about them!

Learning Objectives

Keep these learning objectives in mind as you read to help you identify the most critical information in this chapter.

By the end of this chapter, you should be able to:

1. [LO1](#) Describe the purpose of program evaluation research and five types of questions that program evaluations can seek to address.
2. [LO2](#) Compare and contrast the one-group posttest-only design with the one-group pretest-posttest design.
3. [LO3](#) Describe threats to internal validity for quasi-experimental designs.
4. [LO4](#) Discuss the advantages of having a non-equivalent control group, and compare and contrast the non-equivalent control group design with the non-equivalent control group pretest-posttest design.
5. [LO5](#) Distinguish between the interrupted time series design and control series design.
6. [LO6](#) Describe single case experimental designs and discuss reasons to use these designs.
7. [LO7](#) Compare cross-sectional, longitudinal, and sequential research designs, including the advantages and disadvantages of each design.

Page 185Simple experiments can be powerful but are not always possible. In the basic experimental design described in [Chapter 8](#), participants are randomly assigned to the levels of the independent variable, and a dependent variable is measured. Responses on the dependent variable for the groups are then compared to determine whether the independent variable affected the dependent variable. If all other variables are held constant, differences on the dependent variable can be attributed to the effect of the independent variable. When designed correctly, these simple experiments can be said to possess high internal validity, which allows us to infer that the independent variable caused any observed differences in the dependent variable. If all aspects of a basic experimental design are present, and absent any threats to internal validity, this design is sometimes called a *true experiment*.

However, not all research topics are amenable to the use of a true experiment. This chapter focuses on the different ways that researchers have adapted the basic experimental design for special research circumstances. Single case experimental designs, relying on just a single participant, were developed for instances when only one appropriate participant is available. Quasi-experimental designs are

employed when random assignment for a variable is not possible (e.g., participant variables such as cultural background). Studying changes that occur with age requires yet another set of variations on the basic experimental design. Keep in mind that as these designs depart from the characteristics of a true experiment, they cannot be used to support claims regarding causality.



LO1 Program Evaluation

Researchers frequently investigate applied research questions and evaluate real-world programs or interventions. Applied research and evaluation research can present numerous practical difficulties that prevent researchers from using true experiments. In addition, at times researchers are invited to participate late in the process and are not able to provide input on the best measurement techniques. It can also be the case that budget limitations rule out some forms of data collection (Bamberger, Rugh, Church, & Fort, 2004). These complexities are rather common for research done in the real world, but this research remains essential and needs to be done. In this section, we introduce program evaluation research, which raises many of the issues found in applied contexts. We then turn to focus on a methodological tool that is particularly useful in applied settings: quasi-experimental designs.

[Program evaluation](#) is research on programs that are proposed and implemented to achieve some positive effect on a group of people ([Chapter 1](#)). Such programs may be implemented in schools, work settings, or entire communities. An example is the “At My Best” program designed to reduce obesity by teaching primary school children to make healthy choices about diet and activity. This program is conducted by classroom teachers with supporting materials developed by Physical and Health Education Canada, along with the AstraZeneca pharmaceutical company. Originally piloted in 2008 in 39 schools, it has since spread to over 1,000 schools across Canada.



Think about It!

You may be wondering about the effectiveness of the “At My Best” program. Because research has yet to be published that examines its short-term and long-term impact, it could be worrisome that so many schools are readily adopting it. As you read about program evaluation and quasi-experimental designs in this chapter, consider how you would design a study to evaluate the effectiveness of this program.

Influential social psychologist Donald Campbell (1969) urged a culture of evaluation in which all programs are thoroughly evaluated to determine whether they are effective. Accordingly, the initial focus of evaluation research was outcome evaluation: Did the program result in the positive outcome for which it was designed? As the field of program evaluation has progressed, evaluation research has broadened its scope to many other questions. These can be categorized into five broad types of questions that can guide program evaluations (Rossi, Freeman, & Lipsey, 2004), depicted in [Figure 10.1](#). Note that these questions are depicted as parts of a sequential process (Rossi et al., 2004; cf. Davidson, 2005). Which questions are emphasized depends on the purpose of the evaluation.

Figure 10.1 Types of program evaluation research

Needs Assessment

- Are there problems that need to be addressed in a target population?



Program Theory Assessment

- How will the problems be addressed? Will the proposed program actually address the needs appropriately?



Process Evaluation

- Is the program addressing the needs appropriately? Is it being implemented appropriately?



Outcome Evaluation

- Are the intended outcomes of the program being realized?



Efficiency Assessment

- Is the cost of the program worth the outcomes?



The first type of question is the *evaluation of need*. Needs assessments involve asking whether there are problems that need to be addressed in a target population. For example, what services do homeless people need most? Do repeat juvenile offenders have particular personal and family problems that could be addressed by an intervention program? Nunes (1998) conducted a needs assessment study among Portuguese Canadians. Some of the most important overall needs identified were education (e.g., ensuring adequate education), economics (e.g., addressing high unemployment rates), and integration into Canadian society (e.g., gaining access to social services). Data for a needs assessment may come from surveys, interviews, and existing archival data maintained by public health, criminal justice, and other agencies. Once a need has been established, programs can be planned to address the need.

The second type of question addresses *program theory*. After identifying needs, a program can be designed to address them. The program must be based on valid assumptions about the causes of the problems and a cogent rationale for the best way to address these problems (Rossi et al., 2004). Assessing program theory may involve researchers, service providers, and prospective program clients all collaborating to ensure that the program addresses the target population's needs. Assessing program theory includes articulating the rationale for how members will benefit from the program, including how they will access and use the program's services. This rationale can then be evaluated: Will this program actually reach the target population as intended? Does it have appropriate goals?

Page 187

The third type of question is *process evaluation*, or program monitoring. When the program is under way, the researcher monitors whether it is reaching the target population, whether it is attracting enough clients, and whether the staff is providing the planned services. Sometimes, the staff has not received adequate training or the services are being offered in a location that is undesirable or difficult to find. Overall, the researcher seeks evidence that the program is doing what it is supposed to do. This research is extremely important to avoid concluding that a program is ineffective, when really it is not being implemented properly. Such research may involve questionnaires and interviews, observational studies, and analysis of records kept by program staff.

The fourth question concerns *outcome evaluation*, or impact assessment: Are the intended outcomes of the program being realized? Is the goal—to increase literacy or to provide job skills, for example—being achieved? To determine this, the evaluation researcher must devise a way of measuring the outcome and then

study the impact of the program on the outcome measure. We need to know what participants of the program are like and what they would be like if they had not completed the program. Ideally, a true experiment with random assignment to conditions would be carried out to answer questions about outcomes. However, other research approaches, such as the quasi-experimental and single case designs described in this chapter, can be useful ways of assessing the impact of an intervention program when random assignment is not possible or not ethical.

The final program evaluation question addresses *efficiency assessment*. Once it is shown that a program does have its intended effect, researchers can determine whether the benefits are worth the program's cost. Also, the researchers can determine whether the resources used to implement the program might be put to some better use. Is there a better way to carry out the program?

As you may have noticed, a full program evaluation can be an extensive long-term undertaking, particularly if all of the above questions are to be addressed. Researchers such as Bamberger and colleagues (2004) are developing systematic approaches to respond to the specific challenges that arise when doing evaluation research. One example is what they refer to as "shoestring evaluation," when there are constraints related to time, budget, and data collection options. Program evaluation is one of the viable career options for students majoring in the social sciences. See www.evaluationcanada.ca for information about the professional designation and career opportunities. Next we will consider quasi-experimental designs, which are sometimes incorporated into the program evaluation process.

Quasi-Experimental Designs

Quasi-experimental designs address the need to study independent variables in settings in which true experimental designs are not possible (note that *quasi* is a Latin term meaning “as if”). Quasi-experiments resemble experiments in some ways, but lack important features of true experiments out of necessity, such as control conditions and random assignment to conditions. As a result, they cannot be used to make causal inferences, but rather indicate how variables are related (i.e., associations, similar to correlational studies with measured variables). In addition, they provide a good example of how studies that claim to be using a true experimental design might fall short of that goal, and inadvertently include elements that prevent them from being true experimental designs and therefore do not permit causal inferences. It is very important to distinguish between (1) quasi-experimental designs used intentionally and out of necessity and (2) flawed experiments that purport to be true experiments but are actually quasi-experimental in nature. Quasi-experiments are not bad or poor experiments; in fact, in many circumstances, they are used out of necessity. However, quasi-experiments do help to illustrate the issues that prevent a design from being a true experiment, and thereby help us to evaluate studies that claim to use a true experimental design but in truth fall short of this goal. Page 188

There are many types of quasi-experimental designs, including both between-subjects and within-subjects designs (Campbell, 1969; Campbell & Stanley, 1966; Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002). However, in this text we introduce you to only six of these designs. As you read about each design, compare the design features with the true experimental designs described in [Chapter 8](#) and try to identify what specific limitations emerge for the quasi-experimental design. Learning about quasi-experimental designs will help you identify when an experimental design falls short of a true experiment and possesses some threats to internal validity that might not be initially obvious. [Table 10.1](#) helps you organize and compare all of the designs described in this chapter. We start with the simplest of the quasi-experimental designs: the one-group posttest-only design. Page 189

Table 10.1 Research designs for special circumstances, compared with a true experiment

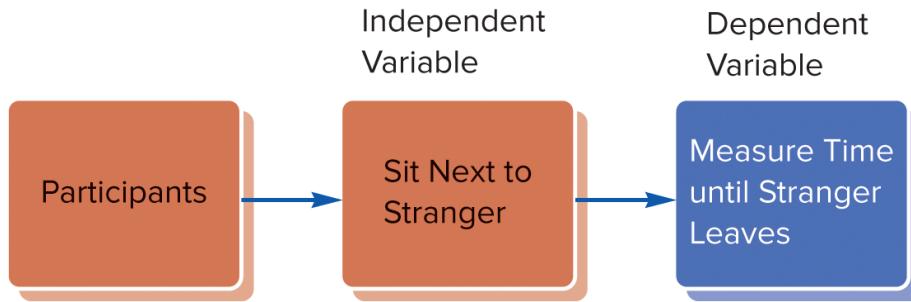
Design Category	Design	Number of Groups/Conditions	Random Assignment to Condition?	Number of Participants	Pretest? Yes	Posttest? Yes
True experiment	Between-subjects	2+	Yes	Many (often 30+ per group)	Sometimes (dependent variable)	Yes
	Reversal ABA(B)	1 (more if re-institute treatment, as in ABAB)	N/A	1	Yes (A, baseline)	Yes (reversal to baseline)
	Multiple baseline	1 (more if multiple settings or behaviours)	No	1 (more if multiple participants)	Yes (A, baseline)	Yes (B, treatment)
Quasi-experiment	One-group posttest-only	1	No	Many (often 30+)	No	Yes

Design Category	Design	Number of Groups/Conditions	Random Assignment to Condition?	Number of Participants	Pretest?	Posttest?
One-group pretest-posttest		1	No	Many (often 30+)	Yes	Yes
Non-equivalent control group		2+	No	Many (often 30+ per group)	No	Yes
Non-equivalent control group pretest-posttest		2+	No	Many (often 30+ per group)	Yes	Yes
Interrupted time series		1	No	Many (often archival data)	Yes, multiple	Yes, often multiple
Control series		2+	No	Many (often archival data)	Yes, multiple	Yes, often multiple
Longitudinal		1	No	Many (often 30+)	Sometimes	Yes, multiple
Cross-Developmental sectional		2+	No	Many (often 30+ per group)	Sometimes	Yes
Sequential		2+	No	Many (often 30+ per group)	Sometimes	Yes, multiple



LO2 One-Group Posttest-Only Design

Suppose you want to investigate whether sitting close to a stranger will cause the stranger to move away. You might try sitting next to a number of strangers and measure the number of seconds that elapse before each leaves. Your design would look like this:



Now suppose that the average amount of time before people leave is 9.6 seconds. Without any sort of comparison, this finding is unfortunately rather uninterpretable. You don't know if people would have stayed longer if you had not sat down, or if people would have stayed for 9.6 seconds anyway. Maybe they liked you and would have left sooner if you had not sat down!

This [one-group posttest-only design](#) lacks a crucial element of true experiments: a control group or other source of comparison. There must be some sort of comparison—and, ideally, random assignment to separate conditions—to enable you to interpret your results. With only one group and one measurement instance, this is not an experiment that will allow us to draw any causal inferences about the effect of an independent variable on a dependent variable, because the results are open to many alternative interpretations (e.g., threats to internal validity, discussed shortly). In other words, it lacks internal validity.

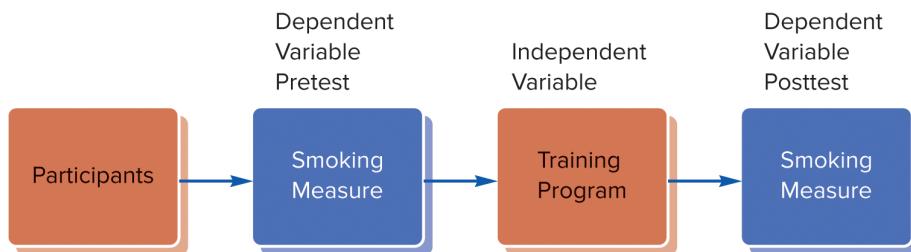
You may see this type of design used as (weak and insufficient) evidence for the effectiveness of programs or advertised products. For example, employees in a company might participate in a four-hour information session on emergency procedures, after which they score an average of 90 percent on a knowledge test. Without any sort of comparison, it would be inappropriate to conclude that the program is successfully educating employees. Remember, this design lacks internal validity. We do not know if the score on the dependent variable would have been equal, lower, or even higher without the program. Likewise, advertisers sometimes make claims about their products based on one-group posttest-only data. It is not enough to know only that children earned A grades after enrolling in an after-school enrichment class (maybe they already earned As), or that people reported high athletic performance after drinking a sport drink (maybe they would have performed even better if they drank water instead). People and companies sometimes report these kinds of results because they do not realize that these data are insufficient evidence, or because they are intentionally trying to mislead consumers. As scientists, we know that we need comparison data to be able to interpret any result. However, there may be some unique circumstances in which a one-group posttest-only design is the only design possible, making it necessary to use this design.

One-Group Pretest-Posttest Design

One way to obtain a comparison (and therefore increase internal validity) is to measure participants before the manipulation (a pretest) and again afterward (a posttest). An index of change from the pretest to the posttest could then be computed. Although this [one-group pretest-posttest design](#) is an improvement over the one-group posttest-only design, it is still not a true experiment and does possess some issues that prevent the kinds of inferences possible with a true experiment.

To illustrate, suppose you wanted to test the hypothesis that a relaxation training program will decrease cigarette smoking. If you were to use the one-group pretest-posttest design, you would select a group of

people who smoke, measure their smoking rate, have them attend relaxation training, and then measure their smoking rate again. Your design would look like this:



If you found a reduction in smoking, you could not assume that the result was due to the relaxation training program based on this design. This design has failed to take into account several potential alternative explanations. These alternative explanations are called threats to internal validity, and include things we will explore in the next section, such as history, maturation, testing, instrument decay, and regression toward the mean. These threats to validity are a shortcoming of quasi-experimental designs, which is why quasi-experiments should only be used when true experiments are not possible. In addition, these threats to validity may also appear in true experiments that are flawed, and so quasi-experiments help us illustrate how to identify problems with a true experiment.



LO3 Threats to Internal Validity

Internal validity refers to our ability to claim, based on a true experiment, that the independent variable causes changes in the dependent variable. We consider something a threat to internal validity when it allows for some other reasonable alternative explanation for changes to the dependent variable, other than the manipulation of the independent variable. When a true experiment is well-designed, it is free from threats to internal validity. Quasi-experiments, because they do not have all the key elements of a true experiment, provide a good demonstration of threats to internal validity. As an example of these threats, recall our discussion of confounding variables ([Chapters 4](#) and [8](#)), which undermine internal validity by providing an alternative explanation for the results that cannot be ruled out. However, there are many other threats to internal validity for true experiments. We will discuss some of the most common threats here, in the context of quasi-experimental designs, because these designs often illustrate these threats by virtue of not being true experiments themselves. These common threats are summarized in [Table 10.2](#).

Table 10.2 Some threats to internal validity

Threat	Summary	How It Undermines Internal Validity
History	Historical event that affects all or most participants (e.g., natural disaster, media event) and that is not of interest to the researcher.	Provides an alternative explanation for change between pretest and posttest. If there are multiple groups and the event affects the groups differently, this becomes a confound and an alternative explanation for group differences.
Maturation	Natural changes to participants' short-term states (e.g., fatigue) or long-term development (e.g., education), not of interest to the researcher.	Provides an alternative explanation for change between pretest and posttest. If there are multiple groups and the groups mature differently, this becomes a confound and therefore an alternative explanation for group differences.
Testing	Simply taking the pretest influences people's responses to the posttest.	Provides an alternative explanation for change between pretest and posttest.
Instrument decay	Characteristics of the measurement instrument changes with repeated use.	Provides an alternative explanation for change between pretest and posttest.
Regression toward the mean	May occur when participants are chosen or groups are divided based on extreme scores on the pretest, because extreme scores tend to become less extreme on repeated measurement.	Provides an alternative explanation for change between pretest and posttest.
Attrition	Participants leave the study.	Differences in attrition between groups may create group differences, even if groups were initially randomly assigned, offering an alternative explanation for differences between groups.
Selection effects	Groups are divided based on any reason other than random assignment.	Pre-existing group differences offer alternative explanation for differences between groups.
Cohort effects	Groups are divided by age. A special type of selection effect.	Instead of differences being due to age, the unique characteristics of a particular cohort offer an alternative explanation for differences between groups.

History

History effects can be caused by virtually any event that occurs during or after the experimental manipulation, after the pretest (if there is one), but before the posttest. In a one-group design, any such event is confounded with the manipulation. Returning to our smoking example, suppose that a famous person dies of lung cancer during the time between the first and second measures. This event, and not the relaxation training, could be responsible for the reduced smoking observed at the posttest (in other words, the historical event becomes an alternative explanation). Consider this real study of an intervention program targeting the well-being of children whose families had recently immigrated to Montreal (Rousseau, Benoit, Lacroix, & Gauthier, 2009). Just after the pretest, the 2004 tsunami devastated many countries throughout the Indian Ocean region. Because the families under study hailed

from affected regions, the researchers had to interpret the data in light of this historical event. The posttest showed a moderate improvement to well-being after the treatment, but it is impossible to know what the effect would have been if the tsunami had not also occurred at that time. It is possible, for example, that the intervention could have improved well-being to an even greater degree, were it not for this historical event. Page 191

Maturation

People change over time. In a brief period, they become bored, fatigued, perhaps wiser, and certainly hungrier. Over a longer period, people develop or recover from illnesses, adults change or start careers, and children become more coordinated and analytical. Changes that occur in participants systematically over time are called *maturation effects*. In the one-group pretest-posttest design, maturation is confounded with the manipulation. Maturation could be a problem in the smoking reduction example if people generally become more concerned about health as they get older and the posttest occurs after years of treatment. Any such time-related factor might result in a change from the pretest to the posttest, unrelated to the manipulation or intervention, thereby offering an alternative explanation for the results. If this happens, you might mistakenly attribute the change to the treatment rather than to maturation. Page 192

Testing

Testing effects occur when simply taking the pretest changes the participant's behaviour (Chapter 8). For example, a measure of smoking might require people to keep a diary in which they note every cigarette they smoked during the day. This process of keeping track of smoking might reduce smoking by making its frequency salient. Thus, rather than any intervention influencing posttest scores, the pretest measurement might be an alternative explanation for the reduction in smoking. In other contexts, taking a pretest may sensitize people to the purpose of the experiment or make them more adept at a skill being tested. These all provide alternative explanations for changes observed at posttest, other than the manipulation or intervention.

Instrument Decay

Sometimes, the basic characteristics of the measurement instrument, or the way participants use it, change over time; this is called *instrument decay*. Instruments can literally deteriorate: Timers fail as they lose battery power, or software programs develop bugs. Another source of instrument decay occurs when human observers are used to measure behaviour. Over time, a person rating behaviour (the *measuring instrument* in this context) may gain skill in rating, become fatigued, or change the standards on which observations are based. In our smoking example, participants might be highly motivated to record all cigarettes smoked at first when the task is new and interesting, but by the end of the study, they may sometimes forget to record a cigarette. If so, instrument decay becomes an alternative explanation for an apparent reduction in smoking.

Regression toward the Mean

Regression toward the mean can occur whenever participants are selected because they score extremely high or low on some variable (also sometimes called *statistical regression*). When they are tested again, their subsequent scores tend to be closer to the mean. Extremely high scores are likely to become lower (closer to the mean), and extremely low scores are likely to become higher (again, closer to the mean). This is a statistical phenomenon that can sometimes be mistaken for changes due to an intervention or

manipulation, and so serve as an alternative explanation for why a difference in scores might be observed.

Consider once again the smoking example. Regression toward the mean would be a problem if participants were selected because they were initially extremely heavy smokers. By choosing people for the program who scored highest on the pretest, the researcher may have selected many participants who were, for whatever reason, smoking much more than their usual amount at the particular time the measure was administered. Simply because of regression to the mean, the heaviest smokers are, on average, likely to be smoking less when their smoking is measured again. If we then compare the overall amount of smoking before and after the intervention, it will appear that people are smoking less as a result of the intervention. Regression toward the mean is thus an alternative explanation for this reduction in smoking, rather than the intervention.^{Page 193}

Statistically speaking, extreme scores are likely to become less extreme over time, simply because they started out so extreme. The concept of regression toward the mean is sometimes challenging to grasp at first, so let's explore another example. Think about your worst grade in a course ever. Now think about your average grade across all your courses. Your average grade is higher than your worst grade. If we happened to select you for a study investigating an achievement intervention based on your worst grade ever, that grade does not represent your typical performance. So regardless of whether our intervention works, your next course grade is probably going to be higher than your worst ever. Extreme scores tend to become less extreme with repeated measurement. Consider what would happen if we had selected research participants only on the basis of their lowest grade, then gave them an intervention, and then measured their next course grade. Even if our intervention had no effect, our results would look like the achievement intervention helped because the next grades would be higher. Instead, however, regression toward the mean is an alternative explanation for this seeming improvement.

The problem of regression toward the mean is rooted in the reliability of our measures. Recall that any measurement score is composed of a true score plus measurement error ([Chapter 5](#)). Because of this measurement error, with repeated measurements, scores will vary around the true score with most scores close to the true score, but some will be higher and some will be lower. Since measurement error is responsible for this variability around the true score, selecting participants based on extreme scores (far from the mean) means that subsequent scores are likely to be different, and also more likely to be closer to the mean (where most scores fall). For example, if you are, on average, an A student, but then get a grade that is extremely far from that average, such as a D, that D grade is likely due to measurement error (e.g., stress from work or family problems), and subsequent grades are unlikely to be close to that D: They will most likely be closer to your average of an A (i.e., close to your true score, where most scores fall). If we select students for an academic intervention based on extreme scores, odds are that subsequent grades are going to be closer to their mean and better than a D grade. (Note that relatively few students have a true score of D or lower, so some of those who get a D grade will have true scores that are higher than this). As a result, we won't be able to tell if the academic intervention had a positive effect beyond regression to the mean. Regression toward the mean is a problem when participants are selected because of extreme scores on a measure that is not completely reliable. Of course, it is worth emphasizing here that all measures include some measurement error, so in practice there is no such thing as a completely reliable measure.

Regression toward the mean can help us explain everyday events as well. Sports commentators often refer to the hex that awaits an athlete who appears on the cover of *Sports Illustrated* magazine. The performances of a number of athletes have dropped considerably after they were the subject of a *Sports Illustrated* cover story. Although these cover stories might cause the lower performance (perhaps the fame results in nervousness and reduced concentration), regression toward the mean is also a likely explanation. An athlete is selected for the cover of the magazine because of exceptionally high

performance. Regression toward the mean teaches us that very high performance is likely a function of measurement error and also likely to be lower on subsequent measurement. If *Sports Illustrated* also did cover stories on athletes who were in a slump and their performance then increased (i.e., creating a comparison group), we could conclude with greater certainty that regression toward the mean is driving the change in performance, rather than some mystical curse.^{Page 194}

Threats to Internal Validity in One-Group Pretest-Posttest and Experimental Designs

Given its susceptibility to the threats to internal validity explained above, is the one-group pretest-posttest design ever used? Sometimes, in applied settings, a comparison group of any kind is impossible to obtain. In the absence of a comparison group, including a pretest provides a comparison within this one group. Recall the evaluation of a program to teach emergency procedures to employees. With a one-group pretest-posttest design, the knowledge test would be given before and after the training session. Measuring knowledge before the training session allows us to rule out one alternative explanation: people's prior knowledge. However, this does not rule out other alternative explanations, and forming a control group would substantially strengthen this design. However, if no control group is possible, replicating the study at other times with other participants can help rule out some threats to internal validity.



Think about It!

Adding a pretest to a single group design rules out the possibility that participants would have exhibited the effect even without the manipulation or intervention. But what other alternative explanations remain? (Hint: Think about the various threats to internal validity for true experiments.)

It is important to note that many of these threats to internal validity can also affect other designs, including true experiments and longitudinal designs. Let us consider how these threats could affect a true experimental design that uses random assignment. One possibility is that the threat affects one group differently than the other. Perhaps an outside event occurs (i.e., history) that only affects one group of participants, but not the other. Even if the threat affects both groups equally, it could still pose a problem for making causal claims. For example, if the dependent measure becomes less reliable between the pretest and the posttest (i.e., instrument decay) for all groups, any genuine effect of the manipulation might be masked by this increase in measurement error. In this situation, researchers would falsely conclude that the manipulation has no effect, when in truth it is an issue of measurement related to instrument decay. For this example, the threat to internal validity does not take the form of an alternative explanation. Instead, the threat is to the experiment's ability to detect a true effect.

Overcoming Threats to Internal Validity

Many of the threats to internal validity can be addressed by the use of an appropriate control group. A control group (that does not receive the treatment) can help rule out the effects of history, regression toward the mean, and so on. For example, if an outside historical event has the same effect on both the treatment and the control groups, any difference between the treatment and control groups on the dependent variable cannot be attributed to the historical event. The best way to create an equivalent group is using random assignment to condition, but that option is not always available in real-world contexts.

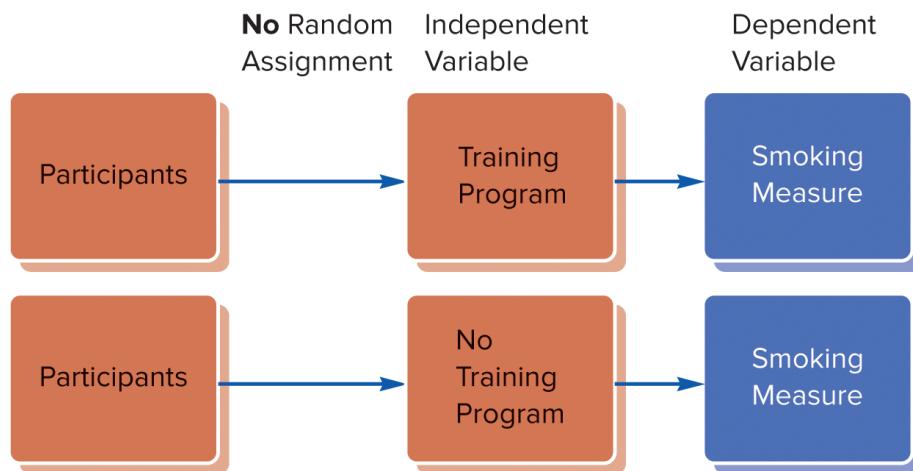
When forming a control group, the participants in the experimental condition and the control condition should be as equivalent as possible. It is not always possible to use random assignment for a between-

subjects design or a within-subjects approach, to ensure that groups are equivalent (e.g., random assignment; [Chapter 8](#)). If participants in the two groups differ before the manipulation, they will probably differ after the manipulation as well, but not necessarily because of your manipulation. The next design illustrates this problem.[Page 195](#)



LO4 Non-equivalent Control Group Design

The [*non-equivalent control group design*](#) has a separate control group, but the participants in the two conditions are not equivalent (e.g., an experimental and a control group). One example is when the participants are not randomly assigned but are instead chosen from naturally pre-existing groups (e.g., students enrolled in the morning section for a course versus the evening section). [*Selection differences*](#) are pre-existing differences between these groups that are confounded with the independent variable, and these differences provide an alternative explanation for the results. If the relaxation training program is studied with a non-equivalent control group design, the design can be diagrammed like this:



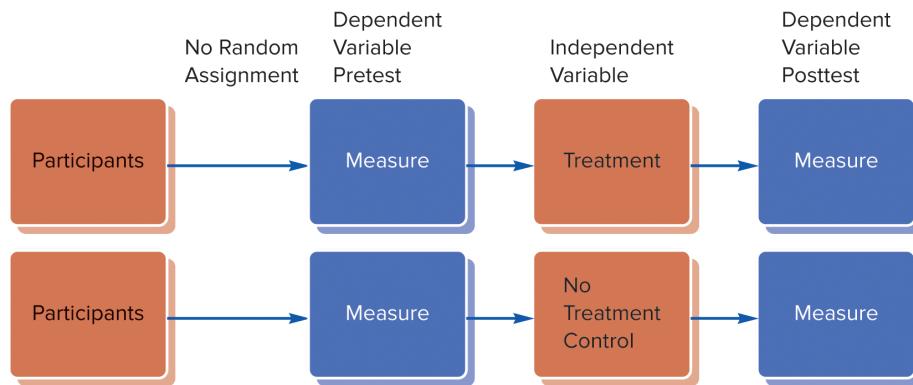
Participants in the first group are given the smoking frequency measure after completing relaxation training, whereas those in the second group do not participate in any program. However, in this design, the researcher does not have any control over which participants are in each group. Suppose, for example, that this study is conducted at a large company. All employees who smoke are invited to participate in the training program, but not everyone signs up for it. The people who do volunteer are in the experimental group, and the people who don't serve as our control group. In this situation, selection differences may arise because those smokers who choose to participate could differ, in some important

way for our constructs of interest, from those who choose not to participate. For example, non-volunteers may smoke far less compared to those who sign up and end up in our experimental condition, or they may be less confident that a program could help them. If these kinds of important selection differences exist, any difference between the groups on the outcome measure may reflect these pre-existing differences rather than any effect of the relaxation training.

Instead of comparing volunteers versus non-volunteers for a program, another option is to compare two similar groups. For example, a researcher might have all smokers in the engineering division of a company participate in the relaxation training program, with smokers who work in the marketing division serve as a control group. In this case, selection differences may still be a problem here, unfortunately. The smokers in the two divisions may have differed in smoking patterns prior to the relaxation program.

Non-equivalent Control Group Pretest-Posttest Design

Adding a pretest to the above design creates a [*non-equivalent control group pretest-posttest design*](#). Although this is still not a true experiment, and therefore not able to support causal inferences, this is one of the most useful quasi-experimental designs. It can be diagrammed as follows:



This is not a true experimental design because assignment to groups is not random, and so the two groups may be different at the beginning. However, we now have the advantage of knowing pretest scores, so we can see whether the groups were the same on the pretest, at least. Moreover, even if the groups are not equivalent, we can look at changes in scores from the pretest to the posttest. If the independent variable has an effect, the experimental group should show a greater change than the control group (Kenny, 1979; for strategies for analyzing these change scores, see Trochim, 2000).

One example of a non-equivalent control group pretest-posttest design looked at how participation in mental health organizations was related to well-being. The participants were all people recovering from mental illness. Some were actively engaged in organizations where they could offer outreach to other people currently experiencing mental illness. These are called consumer/survivor initiatives (CSIs). Others had signed up to participate in CSIs but were not actively engaged in them at the time of study. Researchers asked participants to report on different aspects of their life, and compared how these reports changed over time. After 18 months, active CSI members maintained consistent engagement in employment and education activities, whereas non-active CSI members had reduced engagement in these important activities. Active CSI members also reported an increase in social support over time, and this increase was larger than the increase reported by non-active members. This example is a nice illustration of how random assignment to condition is not always feasible when studying real-world

behaviours. Quasi-experimental designs are the best available tools in these cases, although they fall short of a true experiment. As a result, selection differences still provide a possible alternative explanation for these results. To help address this concern, researchers used a pretest to demonstrate that these groups were similar in terms of demographics and symptom severity at the start of the study.



LO5 Interrupted Time Series Design

The next two designs are similar to the one-group pretest-posttest design and the non-equivalent group pretest-posttest design, respectively. The key addition is that there are multiple pretests and multiple posttests, instead of just one (see [Table 10.1](#)). These designs are commonly used to examine the effects of naturally occurring “manipulations” in society, like the passing of laws. Archival data are often used in these designs ([Chapter 6](#)).Page 197

Perhaps the best way to explain these designs is to begin with an example, starting with a one-group pretest-posttest design and then illustrating how an interrupted time series design can improve upon this design. One group of researchers evaluated Ontario’s crackdown on driving while drunk (Asbridge et al., 2009). At the end of 1996, Ontario passed a law that mandated an immediate 90-day suspended licence for anyone caught driving with a blood alcohol level over the limit. The easiest way to evaluate the effect of this law is to compare the number of driver fatalities in 1996 (before the law) with the number of fatalities in 1997 (after the law). With this design, the researchers found no immediate reduction in the number of driver deaths after the law was passed. This single comparison constitutes a one-group pretest-posttest design with all of that design’s vulnerability to alternative explanations.



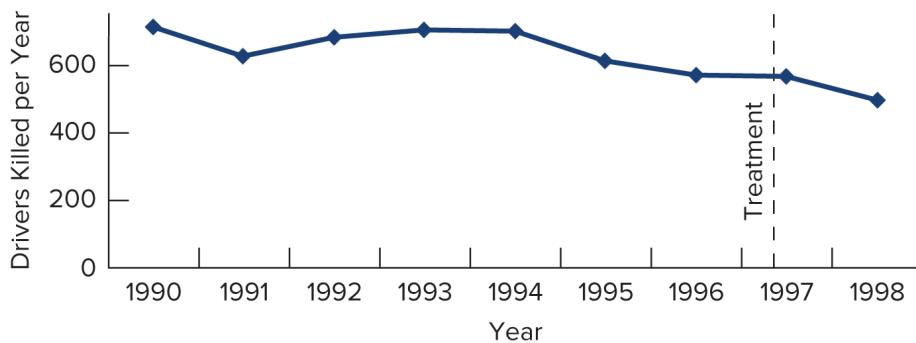
Think about It!

Without reading ahead, re-read the paragraph above and think about one or two ways that you might improve upon this design. I wouldn’t be surprised at all if you came up with the new design we’re about to discuss!

One alternative is to use an [*interrupted time series design*](#), which examines the traffic fatality rates over an extended period of time, both before and after the law was passed. So, if you thought of looking at earlier years prior to the creation of this law, as well as more years afterwards, then you thought of an interrupted time series design. [Figure 10.2](#) shows this information for the years 1990 to 1998. There is a steady downward trend in fatalities after the crackdown, after many years of high fatality rates. Yet, you may notice a remaining problem in interpretation. The year just prior to the intervention, 1996, was already lower than all previous years. It is possible that citizens may have heard of the soon-to-pass law and had begun curbing their drinking and driving behaviour. It is also possible that 1996 was just a

random fluctuation. Still, the data for the years extending before and after the crackdown allow for a less ambiguous interpretation than would be possible, compared to only look at the data for 1996 and 1997.

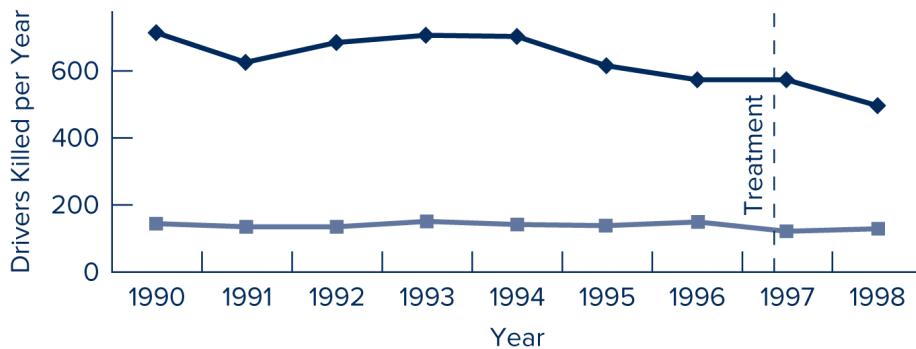
Figure 10.2 Ontario driver fatalities, 1990–1998



Control Series Design

The interrupted time series design can be further improved upon by finding some kind of control group to create a [*control series design*](#). In the case of the Ontario drunk driving law, this was possible because other provinces had not passed any similar law during that time period. So, if you thought of looking at data from other provinces, then you came up with a control series design. [Figure 10.3](#) shows the same data on driver fatalities in Ontario, now adding the fatality rates for Manitoba and New Brunswick (combined) during the same years. The fact that fatality rates in the control provinces remained relatively constant while those in Ontario declined by 14.5 percent led the researchers to conclude that the law had some effect, reducing driver fatalities (Asbridge et al., 2009).

Figure 10.3 Control series design comparing driver fatality rates for Ontario (dark line) and two comparable provinces (light line)



We have now explored some of the various quasi-experimental designs that can be used when circumstances prevent us from using a true experimental design. These designs can provide a valuable opportunity to gain some insight into how variables relate to each other, and they also provide a good illustration for how true experiments might be flawed, or fall short of what is necessary for a true experiment. Next we consider some design options available for researchers interested in studying a single person or changes and growth over time.



Think about It!

Using the table below, think of an example for when each quasi-experimental design might be *necessary*. Importantly, if the situation or context would allow for a true experiment, or allow for a more powerful quasi-experimental design, then try to think of a better example. Consider all sorts of facets of a situation or circumstance that could influence the choice of quasi-experimental design, including cost, the stage of the research (e.g., intervention has already begun, without pretest being included), special populations (e.g., unique workplace contexts, with different limitations with respect to time or population), and so forth. Thinking of topics and situations that are relevant to your own life and experience will help you with this task.

Table 10.3 Situations requiring a quasi-experimental design

Design	Example
One-group posttest-only	
One-group pretest-posttest	
Non-equivalent control group	
Non-equivalent control group pretest-posttest	
Interrupted time series	
Control series	



LO6 Single Case Experimental Designs

Single case experiments were developed so that experiments could be conducted within the context of a case study, with just a single research participant (Barlow, Nock, & Hersen, 2009; Shadish et al., 2002). In a single case experimental design, the participant's behaviour is first measured during a baseline control time period, before any manipulation of a variable or intervention. The manipulation is then introduced during a treatment period, and the participant's behaviour continues to be observed. A change in the participant's behaviour from the baseline to treatment periods offers evidence for the effectiveness of the manipulation.

Much of the early interest in single case designs in psychology came from classic research on reinforcement, pioneered by B. F. Skinner (e.g., Skinner, 1953). Now, these designs are often used in clinical, counselling, educational, and other applied settings (Kazdin, 2001; Morgan & Morgan, 2001). For example, Dolhanty and Greenberg (2009) documented the treatment of a woman battling anorexia nervosa. Her scores on measures of eating disorder and depression severity decreased drastically after she received 18 months of emotion-focused therapy. Because this is not a true experiment, however, we cannot be certain that the treatment caused reduction in symptoms. There could be alternative explanations for the changes observed, other than the experimental treatment.

The single case designs described in the following sections attempt to address these concerns.



Think about It!

What other explanations for the improvement observed might be possible? Try to think of as many possible explanations as you can, and read on to learn more about these alternatives and how researchers try to address them.

Reversal Designs

The basic challenge in single case experiments is how to determine that the treatment—specifically—had an effect on the dependent variable. One method is to demonstrate that the effect can be undone, or reversed, by removing the treatment. A simple *reversal design* takes the following form:

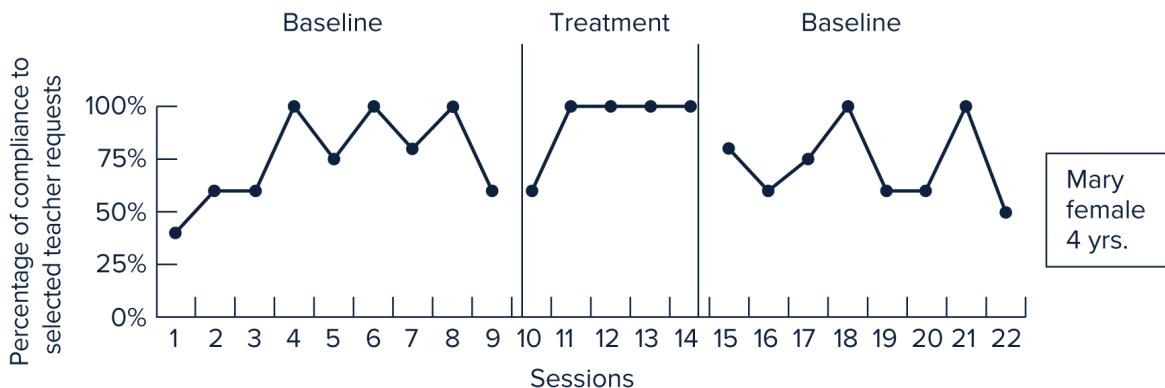
A (baseline period) → B (treatment period) → A (baseline period)

This design, called an ABA design, requires that behaviour be observed and measured during the baseline control period (A), again during the treatment period (B), and also during a second baseline control period (A) after removing the experimental treatment. (This is also called a withdrawal design, in recognition of the fact that the treatment is removed, or withdrawn.) Some treatments produce an immediate change in behaviour, whereas many others require a lengthy treatment period before a change is observed (e.g., as in the case of the woman with anorexia nervosa mentioned above).Page 200

A reversal design was used to measure the effect of an intervention for preschool children in southern Ontario (Levine & Ducharme, 2013). The dependent variable was how often a child complied with their teacher's request, measured every single day during the entire study. This began during the baseline period (A), before any intervention was attempted. During the treatment period (B), a daily five-minute playtime intervention was introduced. For this intervention, the teacher joined the child one-on-one during the class's free-play period, supporting and praising the child. Finally, during a second baseline (A) period, the treatment was discontinued as the measurement of compliance continued. Part of the data are shown in [Figure 10.4](#). The fact that this child's was consistently compliant during the treatment period and became less compliant when the treatment was

withdrawn, can be considered evidence for the treatment's short-term effectiveness.

Figure 10.4 Data from an ABA reversal design



Adapted from Levine, D.G., & Ducharme, J.M. (2013). The effects of a teacher-child play intervention on classroom compliance in young children in child care settings. *Journal of Behavioral Education*, 22, 50–65, Figure 1.

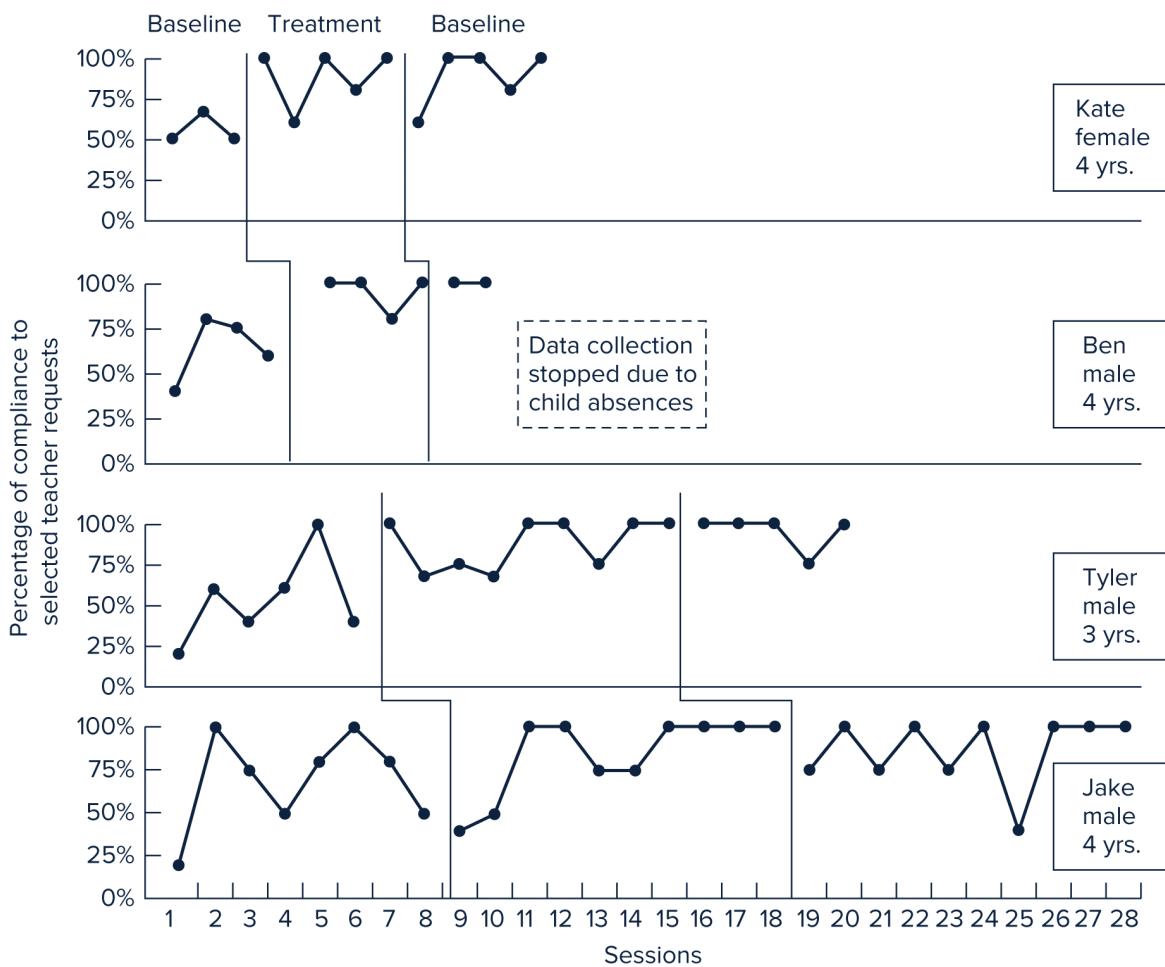


The ABA design can be greatly improved by extending it to an ABAB design, in which the experimental treatment is introduced a second time. This can even be extended to an ABABAB design that allows the treatment to be tested a third time. Adding more reversals can address two issues with interpreting the results from an ABA reversal design. First, a single reversal could be caused by a random fluctuation in the person's behaviour. Perhaps the treatment happened to coincide with some other happy event, such as getting a new puppy, and this is what actually caused the change in symptoms. Coincidental events are less likely to be responsible if the treatment is observed to accompany an improvement two or more times: What are the odds that each time there was another new puppy (or similarly happy event)? The second thing to consider is an ethical concern. It doesn't seem right to end the design with the withdrawal of a treatment that may be very beneficial for the participant (Barlow et al., 2009). An ABAB design not only provides an opportunity to observe a second reversal, it also means that the sequence ends with a treatment rather than its withdrawal.

Multiple Baseline Designs

In a [multiple baseline design](#), researchers look for changes in behaviour after a manipulation is introduced under multiple circumstances (e.g., across different participants or settings). There are several variations of the multiple baseline design (Barlow et al., 2009). In one version, the baseline is measured across participants, with the behaviour of several participants measured over time. The key element of this design is that for each participant, the manipulation is introduced at a different point in time. [Figure 10.5](#) shows data from four additional preschool participants in the play intervention study (Levine & Ducharme, 2013). Note that reports of research using single case experimental designs typically present results separately for each participant, rather than grouping or averaging data.

Figure 10.5 Data from a multiple baseline design across subjects



Adapted from Levine, D.G., & Ducharme, J.M. (2013). The effects of a teacher-child play intervention on classroom compliance in young

children in child care settings. *Journal of Behavioral Education*, 22, 50–65, Figure 2.



Page 201 As you can see in [Figure 10.5](#), introducing the play intervention did not have the exact same effect on compliance across children. However, on average, each child did become more compliant during the intervention phase. Because this change occurred across participants, and the intervention was introduced at a different time for each participant, some alternative explanations can be ruled out. For example, it now becomes unlikely that the result can be attributed to random chance or certain events. By including additional children and starting the intervention at different times for each (i.e., a multiple baseline across participants design), the researchers have made a stronger case for the efficacy of this intervention, compared to using only one child.

In a multiple baseline across behaviours design, several different behaviours of a single participant are measured over time. At different times, an intervention is introduced to target each of these behaviours. For example, a reward system could be introduced to increase the socializing, grooming, and reading behaviours of a child, with each intervention occurring at different times. Observing an improvement in each behaviour when the reward system was applied would be evidence for the effectiveness of the manipulation.

The third variation of the multiple baseline design involves interventions that take place across situations. Here, the same behaviour is measured in different settings, such as at home and at school. Again, a manipulation is introduced at a different time in each setting, with the expectation that a change in the behaviour in each setting will occur only after the manipulation. For example, researchers in Sherbrooke, Quebec, were able to effectively teach a woman with early Alzheimer's disease to find her way around her seniors' residence. These researchers used an ABA design, where (B), the treatment, was applied to multiple locations (e.g., the games room, the laundry room; Provencher, Bier, Audet, & Gagnon, 2008). By demonstrating that the intervention could generalize to different locations in the seniors' residence, they increase our confidence in its efficacy. We now know that its potential is not limited to just one setting or situation.
Page 202

Some multiple baseline designs include a reversal phase and others do not. This might be informed by whether a reversal of some behaviours is impossible or unethical. For example, it would be unethical to reverse a treatment that reduces

dangerous or illegal behaviours, such as kleptomania (i.e., compulsive theft) or alcoholism. Other treatments (e.g., surgery) might produce an irreversible change in behaviour, and so a reversal design is not possible. In such cases, multiple measures over time can be made before and after the manipulation. If the manipulation is effective, a change in behaviour may be observed immediately and the change should continue to be reflected in further measures of the behaviour.

Replications in Single Case Designs

Single case designs have the same limitations as descriptive case studies (e.g., questionable generalizability; [Chapter 6](#)). However, when the results observed with a single participant can be replicated with other participants, the generalizability of these results is enhanced. Consider the full dataset from the play intervention study (Levine & Ducharme, 2013). The treatment was applied to eight children (three girls) from five different schools. For all children, average daily compliance increased during the intervention phase. For six of the eight children, compliance continued after the intervention ceased (e.g., see [Figure 10.5](#)); compliance dropped after the intervention for two children (e.g., see [Figure 10.4](#)). For these two children, compliance was maintained after repeating the intervention (i.e., an ABABA design, not shown). Because the findings were replicated across different children at different schools, the researchers concluded that this simple five-minute intervention can improve compliance among some children.

Single case designs can be especially valuable for applying treatment to help someone in particular improve their behaviour—for example, a parent who is trying to reach a misbehaving child or a clinician who is exploring therapeutic options for a patient. They can even be used to experiment on yourself! Perhaps you’re curious about how a new diet will affect you: A single case experiment might be help you decide whether changing the way you eat will change the way you feel. These designs offer systematic ways to examine hypotheses when studying only one or a few participants is either necessary or desired. Having only one participant has historically hindered statistical analyses, but new techniques for single case designs are being developed (for a review, see Shadish, 2014). Moreover, the results from many single case designs can be combined using meta-analysis to reveal overall patterns ([Chapter 14](#)). For example, a University of Alberta team combined data from 115 single case studies that each studied one to seven participants, for a total of 343 participants (Wang, Parrila, & Cui, 2013). These combined data revealed that social skills interventions can

effectively help people with autism spectrum disorder. In conclusion, despite their limitations, single case designs are a useful, and sometimes necessary, form of research.



Developmental Research Designs

Developmental psychologists often study the ways people change as they age. These topics include changes in reasoning ability as children grow older, the age at which self-awareness develops in young children, or the values people report as they move from young adulthood through to old age. In all cases, the major variable is age. Developmental researchers use two general methods for studying people of different ages: the cross-sectional method and the longitudinal method. You will see that the cross-sectional method shares some similarities with the between-subjects design, whereas the longitudinal method has similarities with the within-subjects design. In this chapter, we will also examine a hybrid approach called the sequential method. These three approaches are illustrated in [Figure 10.6](#).

Figure 10.6 Three designs for developmental research

CROSS-SECTIONAL METHOD

CROSS-SECTIONAL METHOD

	Year of Birth (Cohort)	Time 1: 2020
Group 1:	1965	55 years old
Group 2:	1960	60 years old
Group 3:	1955	65 years old

LONGITUDINAL METHOD

	Year of Birth (Cohort)	Time 1: 2020	Time 2: 2025	Time 3: 2030
Group 1:	1965	55 years old →	60 years old →	65 years old

SEQUENTIAL METHOD

	Year of Birth (Cohort)	Time 1: 2020	Time 2: 2025	Time 3: 2030
Group 1:	1965	55 years old →	60 years old →	65 years old
Group 2:	1955	65 years old →	70 years old →	75 years old

Longitudinal Method

In the longitudinal method, the same group of people is observed at different points in time as they grow older. Some longitudinal studies study people over only a few years. For example, an eight-year study of Swedish children demonstrated positive effects of day care (Broberg, Wessels, Lamb, & Hwang, 1997). Other studies span much larger time frames. One famous longitudinal study is the Terman Life-Cycle Study, which was

begun by Stanford psychologist Lewis Terman in 1921. Terman studied over 1,500 California schoolchildren who had intelligence scores of at least 135. The participants, who called themselves “Termites,” were initially measured on numerous aspects of their cognitive and social development in 1921 and 1922. Terman and his colleagues continued studying the Termites during their childhood and adolescence and throughout their adult lives (cf., Terman, 1925; Terman & Oden, 1947, 1959). To this day, Terman’s successors at Stanford continue to track the Termites until each one dies. The study has provided a rich description of the lives of highly intelligent people and offered evidence disputing many negative stereotypes of high intelligence. For example, the Termites were quite well-adjusted, both socially and emotionally. These data have been archived for use by other researchers, such as Friedman and colleagues (1995), who used the Terman data to study factors associated with age at death. One intriguing finding was that the personality trait conscientiousness predicted longevity. More recently, researchers have found that sleeping more or less than the group’s average during childhood predicted earlier deaths, but only for the male Termites (Duggan, Reynolds, Kern, & Friedman, 2014).Page 204

A longitudinal study on aging and Alzheimer’s disease called the Nun Study illustrates a different approach (Snowden, 1997). In 1991, all members of a religious order born prior to 1917 were asked to participate by providing access to their archived records as well as complete annual measures taken over the course of the study. The sample consisted of 678 women with a mean age of 83. One fascinating finding from this study was based on autobiographies that all women wrote in 1930 (Donner, Snowden, & Friesen, 2001). The researchers devised a coding system to measure positive emotional content in these autobiographies ([Chapter 6](#)), and greater expression of positive emotions was found to predict longevity.

Cross-Sectional Method

In contrast to the longitudinal method, which follows people over time, the [*cross-sectional method*](#) involves studying people of different ages at a single point in time. Suppose you are interested in examining how the ability to learn new software changes as people grow older. Using the cross-sectional method, you might study people who are currently 20, 30,

40, and 50 years of age. The participants in your study would be given the same software learning task, and you would compare the different age groups on their performance.

Comparing Longitudinal and Cross-Sectional Methods

The cross-sectional method is much more common than the longitudinal method, primarily because it is less expensive to conduct and yields immediate results. It would take 30 years to study the same group of people from age 20 to 50 with a longitudinal design, but with a cross-sectional design these comparisons can be obtained relatively quickly.

There are, however, some disadvantages to cross-sectional designs, the most important of which is the fact that the researcher cannot be sure that differences among age groups are due only to age. For a cross-sectional design, developmental change is not observed directly within the one group of people, but is based on comparisons among different cohorts. You can think of a cohort as a group of people born at about the same time, exposed to the same events in a society, and influenced by the same demographic trends (e.g., divorce rates and family size). People who are different ages at this moment all grew up in different eras, and this might explain any differences between age groups rather than aging or development itself. If you think about the hairstyles of people you know who are in their 20s, 40s, and 60s, this might help you imagine how growing up in a particular era might shape a person. More crucially, differences among cohorts reflect different economic and political conditions in society, different musical trends and available technologies, different educational systems, and different child-rearing practices. In a cross-sectional study, a difference among different age groups may well reflect developmental age changes, but they could also be the product of *cohort effects* (Schaie, 1986). Since we cannot separate out the effects of cohort from age in a cross-sectional design, these two things are confounded.

To illustrate this issue, let's consider the hypothetical study on learning to use new software. Suppose your results showed that age is associated with

lower ability, such that the 50-year-olds are worse at learning new software than the 40-year-olds, and so on. Should you conclude that the ability to learn to use a new program decreases with age? That may be an accurate conclusion. Alternatively, the differences could be due to a cohort effect. Perhaps the older people in your sample had less experience with software and computers while growing up. We would not expect today's 20-year-olds to have as much difficulty learning to use new software in 40 years, compared to today's 60-year-old person who had far less exposure to computers when they were in their 20s (40 years ago). The key point here is that the cross-sectional method confounds cohort effects with the variable of interest: age. Cohort effects are most likely to be a problem when the researcher is examining age effects across a wide range of ages (e.g., adolescents through older adults). The life experiences of adolescents versus retired adults are likely to be much more different than the life experiences of people in their mid-teens versus early twenties, for example.

Page 205

The only way to conclusively study changes that occur as people grow older is to use a longitudinal design, following people as they age. If a researcher wants to study how the home environment of children at age 5 is related to school achievement at age 13, data from a longitudinal study provides the best evidence. An alternative to a longitudinal design in this case would be to study 13-year-olds and ask them or their parents about their earlier home environment. This is known as a retrospective approach, and it introduces other challenges, such as potential difficulty remembering past events accurately.

The major limitations of a longitudinal approach are its difficulty and expense. It is hard to keep track of many people over a number of years. In addition, over the course of a longitudinal study, people may move, die, or lose interest in participating. In other words, there is a great risk of high participant attrition ([Chapter 8](#)). Researchers who conduct longitudinal studies become adept at convincing people to continue, often travelling to the participant to collect data or scouring social networking sites to maintain connections over time (Mychasiuk & Benzies, 2011). To account for attrition when analyzing the data, researchers must compare the scores of people who drop out with those who stay. If those who stay seem quite

different from those who leave, then we might become concerned about whether there is something unique and unidentified about the final sample being analyzed. A researcher shouldn't embark on a longitudinal study without considerable resources and a great deal of patience and energy.

Sequential Method

The *sequential method* blends both the longitudinal and cross-sectional methods. The first goal of the sequential study depicted in [Figure 10.6](#) is to compare 55- and 65-year-olds. This is achieved by using a cross-sectional method, studying different groups of 55- and 65-year-olds. All participants are then studied using the longitudinal method, with each person tested at least one more time in the future (e.g., five years later). Using this method may take fewer years to complete than a traditional longitudinal study, and cross-sectional data are ready to analyze while longitudinal data are being collected.

We have now described most of the major approaches to designing research. Earlier in this text we examined true experimental designs ([Chapters 4](#) and [8](#)), as well as various methods typically used in non-experimental designs (correlations and surveys in [Chapters 4](#) and [7](#), observational methods in [Chapter 6](#)). The current chapter addressed a variety of options available when a researcher wishes to conduct a true experiment, but practical circumstances prevent it (e.g., no access to a control group, unable to randomly assign people to conditions). In doing so, it also illustrates potential flaws in experiments that purport to be true experiments, but fall short of the standard. In the next chapter we return to the true experimental design, introducing more complex versions of this powerful method. Following this, we introduce the foundations of data analysis using statistics.[Page 206](#)



Illustrative Article: A Longitudinal Study

In what ways do young adolescents change as they mature into young adults? What happens when typical teen behaviour happens a bit too early?

Allen, Schad, Oudekerk, and Chango (2014) examined data from a longitudinal study in which adolescents were measured at 13, 14, and 15 years old, and then again at 23 years old. They focused specifically on early adolescent “pseudomature” behaviours such as vandalism or shoplifting, romantic involvement, and focus on physical appearance. They were interested in what happens to these “cool” kids, who enjoy high status among peers in early adolescence, ten years later.

First, acquire and read the article:

- Allen, J. P., Schad, M. M., Oudekerk, B., & Chango, J. (2014). What ever happened to the “cool” kids? Long-term sequelae of early adolescent pseudomature behavior. *Child Development*, 85, 1866–1880. doi:10.1111/cdev.12250

Then, after reading the article, consider the following:

1. How did the researchers operationalize pseudomature behaviour? What is your opinion of their operationalization?
2. The researchers note, “Pseudomature behavior has been cross-sectionally associated with greater popularity among peers” (p. 1867). They use this finding to support their method. How is their approach different from a cross-sectional approach? What are the strengths and weaknesses of their approach?
3. The authors had five hypotheses. Choose one, describe how the authors arrived at the hypothesis, and then describe the actual results related to the hypothesis.
4. Interpret table 5 in the article: Find the section of the article that is associated with table 5 and compare the written results with the information in the table.

5. Interpret figure 1 in the article: What does it tell you about popularity change over time?
6. What if the researchers had continued to follow these same participants? Do you think that the conclusions about long-term implications of early pseudomature behaviour would be the same when the participants were 33 years old? 43?
7. The authors identified a variety of limitations of their study. Choose one and describe how you would “fix” it.

Study Terms

Test yourself! Define and generate an example of each of these key terms.

- baseline (p. 199)
- cohort effects (p. 204)
- control series design (p. 197)
- cross-sectional method (p. 204)
- history effects (p. 204)
- instrument decay (p. 192)
- interrupted time series design (p. 197)
- longitudinal method (p. 203)
- maturity effects (p. 192)
- multiple baseline design (p. 200)Page 207
- non-equivalent control group design (p. 195)
- non-equivalent control group pretest-posttest design (p. 195)
- one-group posttest-only design (p. 189).
- one-group pretest-posttest design (p. 190)
- program evaluation (p. 185)
- quasi-experimental designs (p. 187)
- regression toward the mean (p. 192)

- reversal design (p. 199).
- selection differences (p. 195).
- sequential method (p. 205).
- single case experimental design (p. 199).
- testing effects (p. 192).

Review Questions

Test yourself on this chapter's learning objectives. Can you answer each of these questions?

1. What is a reversal design? How is an ABAB design superior to an ABA design?
2. What is meant by baseline in a single case design?
3. What is a multiple baseline design? Why is it used? Distinguish among multiple baseline designs across participants, across behaviours, and across situations.
4. List five types of program evaluation research questions. What research goals does each address?
5. Why might a researcher use a quasi-experimental design rather than a true experimental design?
6. Why does adding a control group eliminate some problems associated with the one-group pretest-posttest design?
7. Describe the threats to internal validity discussed in this chapter. How do they threaten internal validity?
8. Describe the non-equivalent control group pretest-posttest design. What makes this design a quasi-experiment rather than a true experiment?
9. Contrast the interrupted time series from the control series designs. Which design provides stronger evidence? Why?
10. Compare the features, advantages, and disadvantages of longitudinal, cross-sectional, and sequential methods. In which of these designs are cohort effects problematic? Explain your rationale.

Deepen Your Understanding

Develop your mastery of these concepts by considering these application questions. Compare your responses with those from other people in your study group.

1. Your dog gets lonely while you are at work and consequently engages in destructive activities such as pulling down curtains or strewing garbage all over the floor. You decide that playing music while you are gone might help. How might you determine whether this “treatment” is effective?
2. The captain of each precinct of a metropolitan police service selected two officers to participate in a program designed to reduce prejudice by increasing sensitivity to racial and ethnic group differences. The training program took place every Friday morning for three months. At the first and last meetings, the officers completed a measure of prejudice. To assess the program’s effectiveness, the average prejudice score at the first meeting was compared with the average score at the last meeting. It turns out that the average score was in fact lower following the training program. What type of design is this? What specific problems arise if you try to conclude that the training program was responsible for the reduction in prejudice?Page 208
3. Many elementary schools have implemented a daily silent reading period during which students, faculty, and staff spend 20 minutes silently reading a book of their choice. Advocates of this policy claim that the activity encourages pleasure reading outside the required silent reading time. Design a non-equivalent control group pretest-posttest quasi-experiment to test this claim. Include a dependent measure and provide a rationale for choosing it. Is a true experiment possible in this case? Why or why not?
4. Dr. Cardenas studied political attitudes among different groups of 20-, 40-, and 60-year-olds. Political attitudes were found to be most

conservative in the age-60 group and least conservative in the age-20 group.

1. What type of method was used?
2. Can you confidently conclude that people become more politically conservative as they get older? Why or why not?
3. Propose alternative ways of studying this topic.

Complex Experimental Designs



©Alexander Semenov/Getty Images

Even seemingly simple creatures like jellyfish are actually enormously complex. Let's take those basic experimental principles and see how they apply to more complex experimental designs.

Learning Objectives

Keep these learning objectives in mind as you read to help you identify the most critical information in this chapter.

By the end of this chapter, you should be able to:

1. [LO1](#) Define a *factorial design* and discuss reasons why a researcher would use this design.
2. [LO2](#) Describe what information is provided by main effects and interaction effects in a factorial design.
3. [LO3](#) Interpret a graph depicting results from a 2×2 experimental design. Use this graph to estimate whether there is one main effect, two main effects, and/or an interaction effect.
4. [LO4](#) Discuss the role of simple main effects in interpreting interactions.
5. [LO5](#) Describe an independent variable by participant variable (IV \times PV) design.
6. [LO6](#) Compare the assignment of participants in a between-subjects design, a within-subjects design, and a mixed factorial design.

Page 210Complex questions sometimes require complex experimental designs. So far we have focused primarily on the simplest true experiment, in which one independent variable is manipulated with two levels, and one dependent variable is measured. Our discussions of how to conduct a study ([Chapter 9](#)) and quasi-experimental designs ([Chapter 10](#)) have illustrated just how difficult it can be to craft a good experiment, with some situations making it impossible to use an experimental design and necessitating a quasi-experiment. In addition, researchers often investigate problems that demand more complicated designs than the simple experiment we have described thus far, and it is to these complex experimental designs that we turn to in this chapter. The fundamental aspects of all experimental research, such as internal validity and procedures for assigning participants to conditions, still hold true for these more complex designs.

An Independent Variable with More Than Two Levels

In the simplest experimental design, the independent variable has only two *levels* (also called groups or conditions). However, a researcher might want to have three or more levels of the independent variable for several reasons. Researchers are frequently interested in comparing more than two groups. For example, in a study examining whether virtual reality (VR) can be used to reduce stress, stressed-out participants were randomly assigned to explore one of three virtual reality environments: nature, a city, or a control condition consisting of simple geometric shapes (Valtchanov & Ellard, 2010). People who explored a natural environment (virtually) experienced the greatest decrease in stress, compared to participants in both the city and shape conditions. In this example, adding the third condition (shape) allowed the researchers to add an additional control condition that could potentially provide greater clarity. For example, if the participants in both the nature and the city conditions experienced a reduction in stress, then the shape condition might rule out that VR exposure alone reduces stress (indicating that it might have to do with exploring realistic environments in VR).

Another reason to have more than three conditions is that a design with only two levels of the independent variable may not provide enough information about the relationship between the independent and dependent variables. Consider a study where two groups of people are compared with respect to performance on a motor task. One group is promised \$40 for excellent performance, whereas the control group is not promised a reward. The solid line in [Figure 11.1](#) depicts the results of this hypothetical study: We see near perfect performance for the \$40 condition and much worse performance for the control condition (about 50 percent correct). Because there are only two levels of the independent variable, the relationship between reward and performance can only be described with a straight line. We do not know if the true relationship between these variables is described by this straight line (i.e., a linear relationship). In order to examine whether this relationship is truly linear, we would need to examine different levels of reward, between \$0 and \$40. The broken line in [Figure 11.1](#) illustrates the possible results when \$10, \$20, and \$30 are also included as levels of the independent variable.

Figure 11.1 Results of a hypothetical experiment: positive linear versus positive non-linear functions



This result is a more nuanced description of the relationship between reward (i.e., the independent variable) and performance (i.e., the dependent variable). Now we can see that adding a reward increases performance, but this relationship begins to flatten out or become weaker as the amount increases. Offering a cash reward is very effective in increasing performance up to a point (i.e., around \$20), after which only modest increases in performance accompany increases in reward. Thus, the relationship is a non-linear positive relationship rather than a strictly linear relationship ([Chapter 4](#)). Knowing this would be very useful for deciding how much money you need to pay employees or participants to do a similar task. If we only had two levels of the independent variable, we would have missed this information.

Recall from [Chapter 4](#) that variables are sometimes related in a curvilinear fashion: direction of relationship changes across values of the variable. [Figure](#)

[11.2](#) shows an example of a type of curvilinear relationship called an inverted-U, so named because the curve has an upside-down U shape. An experiment with only two conditions cannot detect curvilinear relationships like an inverted-U.

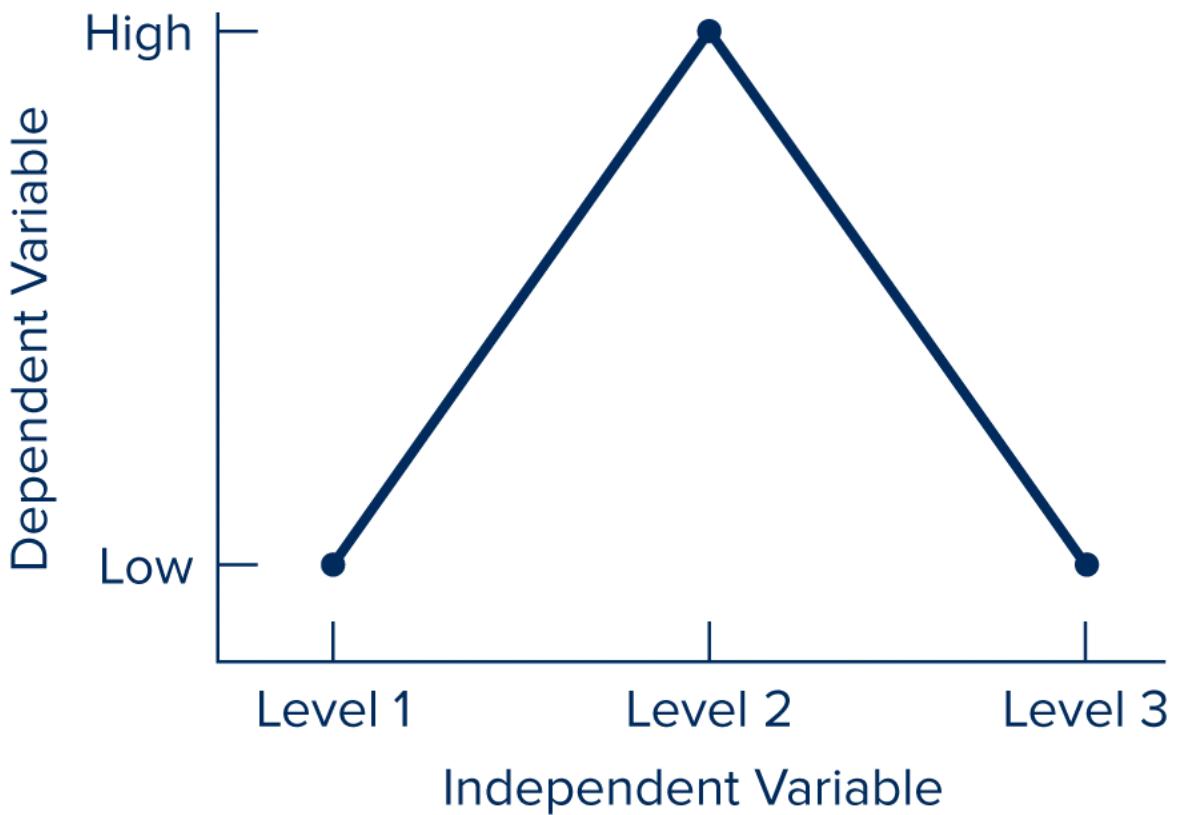


Think about It!

Imagine a study in which the true nature of the relationship between the independent and dependent variable looks like the results depicted in [Figure 11.2](#). Now imagine that you ran a study with only two conditions: Level 1 and Level 2. What sort of conclusion would you draw about how these two variables are related? If you ran a study with only Level 1 and Level 3, what sort of conclusion would you draw?

If a curvilinear relationship is predicted, at least three levels must be used. Many such curvilinear relationships exist in psychology. The relationship between fearful arousal and attitude change is one example. Increasing the amount of fear aroused by a persuasive message increases attitude change up to a moderate level of fear, but further increases in fear actually reduce attitude change.

Figure 11.2 Curvilinear relationship



Note: At least three levels of the independent variable are required to show curvilinear relationships.





LO1 An Experiment with More Than One Independent Variable: Factorial Designs

Our simple experiment had one independent variable (e.g., reward) with two levels (e.g., no reward versus \$40). In addition to adding additional levels of an independent variable (e.g., \$10 and \$20 rewards), you can also add additional independent variables to an experiment (e.g., task difficulty). Adding more independent variables to an experiment brings it closer to real-world conditions, in which many influences (i.e., independent variables) interact with one another to produce some behaviour or effect.



Think about It!

How are you feeling today, right at this moment? Now think of all the different reasons why you feel the way you do (e.g., what you ate recently, when you last ate, what you're going to eat later). This illustrates just how complex real-world phenomena are, and how many things influence us to create a particular behaviour or experience.

Recall the hypothetical crowding experiment that was described in [Chapter 8](#). For experimental control, participants in all conditions were run with the same number of people in the room. Another option is to study whether the number of people influences test performance—to use this as an additional independent variable. In this case, your experiment would now have two independent variables: crowding and number of people present. Having more than one independent variable allows you to examine how these two variables influences the dependent variable, and also how they might interact. These designs are known as factorial designs.

A [*factorial design*](#) has more than one independent variable (also known as a factor), and with each independent variable (IV) also having more than one level. The simplest factorial design has two independent variables, each with two levels. We typically describe a factorial design using a particular shorthand that looks like this:

- Number of levels of first IV \times Number of levels of second IV

For two independent variables, each with two levels, this would be described as a “ 2×2 factorial design” (read as a “two by two design”).

An experiment by McFerran and colleagues (2010) nicely illustrates a 2×2 factorial design. These researchers studied the effects of body type on others’ food consumption. Participants were told that the study was about experiences viewing movies, and they were offered a snack while viewing these movies. All participants were run in pairs, but the other “participant” was actually a confederate. One independent variable was the amount of snack food the confederate selected (30 candies or 2 candies). The other independent variable was the confederate’s body size, which also had two levels: thin versus obese. Researchers had a special suit created by an award-winning costume designer, which increased the confederate’s natural size from a size 00 to a size 16, adding approximately 80 pounds (34 kg) to her frame. (If you’d like to see pictures of this amazing transformation, use the reference to find this article.) The dependent variable was the number of candies the real participant ate.

This 2×2 design results in four experimental conditions (calculated by multiplying two by two), each with a different combination of the two independent variables ([Table 11.1](#)). Participants were randomly assigned to experience one of these four combinations: (1) thin confederate—30 candies, (2) thin confederate—2 candies, (3) obese confederate—30 candies, and (4) obese

confederate—2 candies. We will use this study throughout this chapter to illustrate key concepts in factorial designs. Page 213

Table 11.1 2×2 factorial design: Results of the body type and food choices study

Confederate food selection (Independent Variable A)	Confederate body type (Independent Variable B)		Marginal means (shows main effect of A)
	Thin	Obese	
30 candies	9.82	6.25	8.04
2 candies	3.20	4.26	3.73
Marginal means (shows main effect of B)	6.51 5.26		



LO2 Interpreting Factorial Designs

Factorial designs yield two distinct kinds of information. The first is information about the effect of each independent variable, taken by itself: the main effect of each independent variable. In a design with two independent variables, there are two main effects, one for each independent variable. The second type of information is called an interaction. If there is an interaction between two independent variables, the way that one independent variable affects the dependent variable depends on the level of the other variable. Interactions are a

very valuable source of information that does not exist in simple experiments with only one independent variable.

To illustrate main effects and interactions, we can look at the results of the study on how others influence our food choices (McFerran et al., 2010). [Table 11.1](#) illustrates a common method of presenting results for a factorial design. The number in each *cell* of the table represents the mean number of candies people ate. (The *mean* is the formal term for “arithmetic average”; [Chapter 12](#).) The mean number of candies eaten for participants who were in the thin confederate–30 candies condition can be found in the corresponding cell of the table: It is 9.82. As we explore main effects and interactions, it may be helpful to refer back to [Table 11.1](#) each time a variable or value is discussed.

Main Effects

A main effect is the effect that each independent variable has on the dependent variable by itself. The main effect of independent variable A (e.g., confederate’s food choice in the example above) captures its overall effect on the dependent variable (e.g., participants’ own candy consumption). Similarly, the main effect of independent variable B captures the effect of this independent variable (e.g., confederate body type) on the dependent variable.

Essentially, the main effect pretends that the other independent variable doesn’t exist in the experiment. Consider independent variable B: Is there a relationship between confederate body type and candy consumption? We can find out by looking at the *marginal mean* in the thin and obese conditions. These means are shown in the bottom row of [Table 11.1](#). (This bottom row, along with the rightmost column, are called the margins of the table, hence marginal means.) The overall amount of candy eaten by participants in the thin confederate conditions (regardless of how much she ate) is 6.51, and the amount of candy eaten in the obese confederate conditions (regardless of how much she ate) is 5.26. Note that the marginal mean of 6.51 in the thin confederate condition is the average of 9.82 in the thin confederate–30 candies group and 3.20 in the thin confederate–2 candies group. (This calculation assumes equal numbers of participants in each group.) You can see that overall, people ate more candy when in the company of someone who is thin rather than obese.



TRY IT OUT!

What about the main effect for the other independent variable: confederate food selection? Look at [Table 11.1](#) and find the marginal mean for the 30 candies conditions (regardless of confederate body type). What about the marginal mean in the 2 candies conditions (regardless of confederate body type)? Do people eat more or less candy when in the company of someone who takes 30 rather than 2 candies for him or herself?

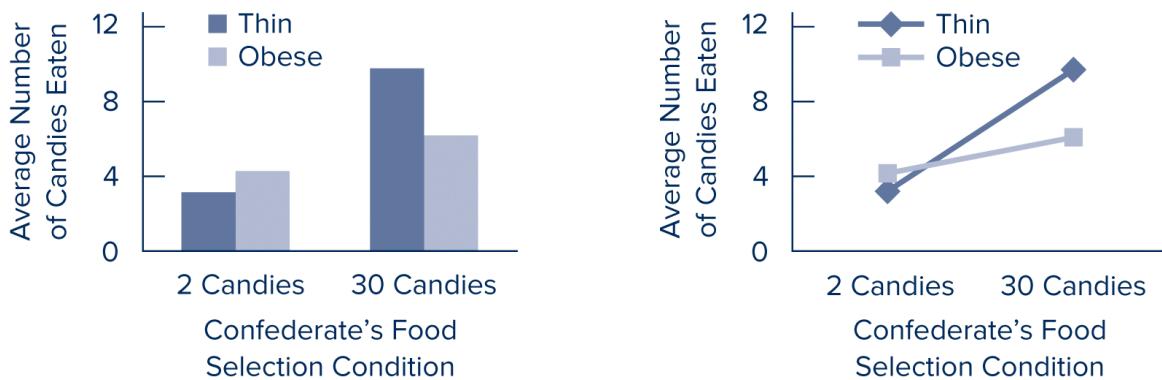
Interactions

The two main effects tell us that overall, people eat more candy when their companion takes lots of candy (rather than a little) and when their companion is thin rather than obese. There is also the possibility that an interaction exists. If so, interpreting the interaction is most important, as it indicates that the main effects must be qualified. In this context, the term qualified refers to the fact that the main effects are conditional or contingent on something else. An interaction between independent variables indicates that the effect of one independent variable varies at different levels of the other independent variable.

Examine [Table 11.1](#) to try to find an interaction. Look at the vertical columns to see that the effect of confederate food selection is different depending on whether she is thin or obese. When a person's eating companion is thin, people eat more candy when she takes many rather than few candies: 9.82 in the 30 candies condition versus 3.20 in the 2 candies condition. However, when the eating companion is obese, the amount of food she takes has a much smaller effect: 6.25 for 30 candies versus 4.26 for 2 candies. Thus, the relationship between the eating companion's body type and food consumption can be understood only by considering both independent variables simultaneously. To know how eating is affected, we must consider the companion's body type *and* whether she takes a lot of food or only a little.

Sometimes it helps to see interactions when the means for all conditions are presented in a graph. So you can compare graph styles, [Figure 11.3](#) shows both a bar graph and a line graph of the same results that are displayed in [Table 11.1](#). Note that all four cell means have been graphed. In the bar graph, two bars (on the left) compare the body types in the 2 candies condition, with the same comparison shown for the 30 candies condition on the right. You can see that an eating companion's body type does not influence candy consumption much when she eats only 2 candies, but when she eats 30 candies her body type has a greater influence. The same interaction can be seen in the line graph on the right.

Figure 11.3 Interaction between companion's body type and food selection on amount eaten, as a bar graph (L) and line graph (R)



Page 215 You probably use the concept of interaction all the time without realizing it. When we say “it depends,” we are usually indicating that some sort of interaction is operating: The result depends on, or is conditional or contingent on, something else. Suppose, for example, that a friend asks you if you want to go to a movie. The likelihood that you will go may reflect an interaction between two variables, such as (1) Is an exam coming up? and (2) Who stars in the movie? If there is an exam coming up, you won’t go under any circumstance. If you do not have an exam to worry about, you may be more likely to go if your favourite actor is in the movie, but less likely to go if your least favourite actor is in the movie. The variables exam and actor interact to affect the likelihood you will go to the movie.



TRY IT OUT!

For practice, try graphing the movie example in the same way as in [Figure 11.3](#). The dependent variable (going to the movie) is always placed on the vertical axis, known as the y -axis. It could have values ranging from 0 (absolutely will not go) to 100 (absolutely will go). The levels of one independent variable (e.g., upcoming exam) are placed on the horizontal axis, or x -axis. Bars are then drawn to represent each of the levels of the other independent variable (i.e., favourite or least favourite actor).

Interactions Illuminate Moderator Variables

In many studies, interactions are discussed in terms of a *moderator variable*. A moderator variable influences the relationship between two other variables (Baron & Kenny, 1986). In the study by McFerran and colleagues (2010), the main effect of a companion's food selection can be stated as, "People eat more food when an eating companion takes more rather than less food to eat." However, because we have an interaction, we must then make a qualifying statement that the companion's body type influences (or moderates) this relationship: "People eat more food when an eating companion overindulges only when the companion is thin; when the companion takes little food, the amount people eat is not influenced by the companion's body type." The body type variable is a moderator variable because it moderates (i.e., changes or influences) the relationship between the other variables. Moderator variables may be aspects of the situation, as in this food study, or they may be aspects of the participants (discussed [later](#)).



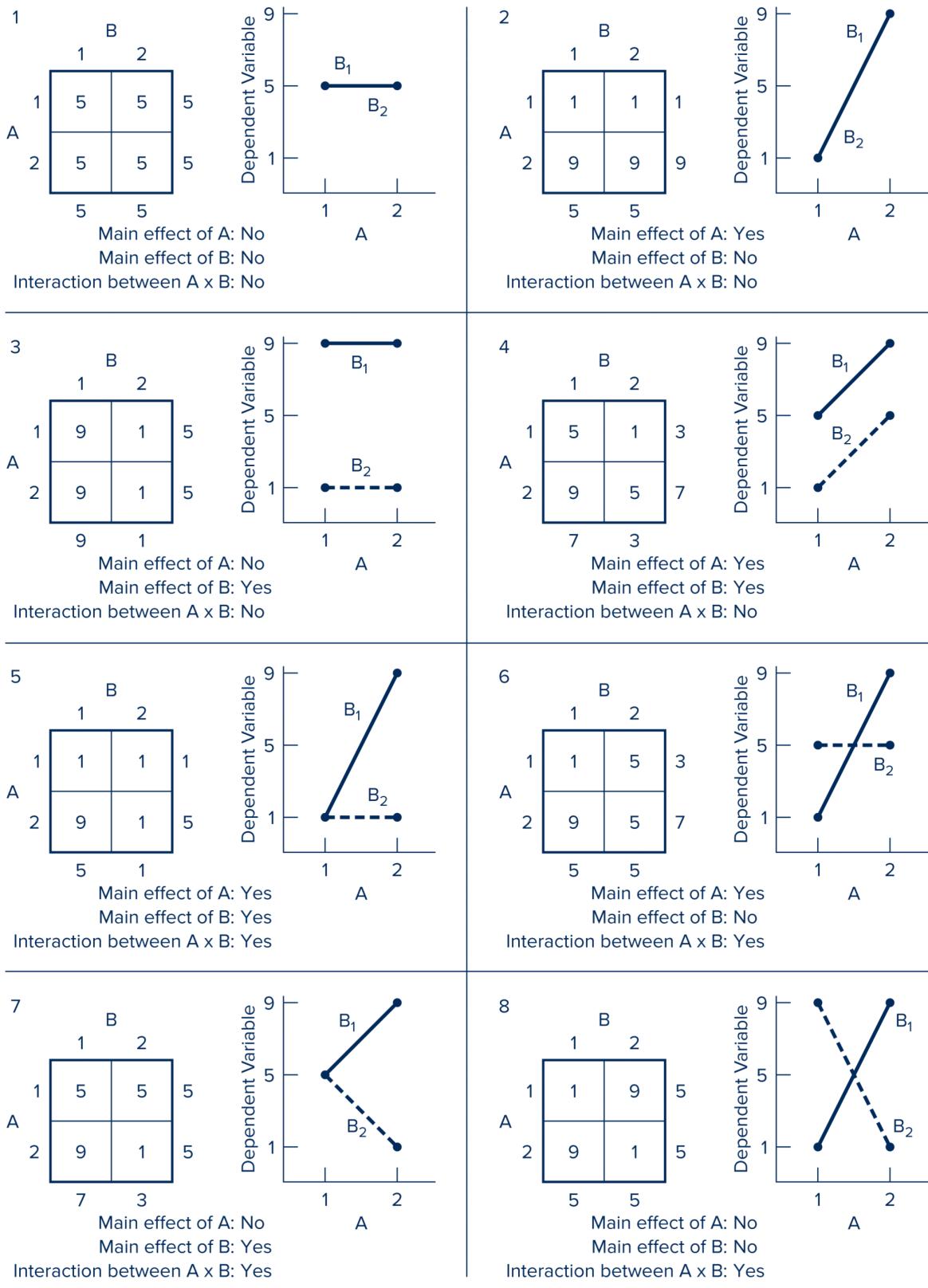
LO3 Depicting Possible Outcomes of a 2×2 Factorial Design Using Tables and Graphs

Recall that a 2×2 factorial design has two independent variables, each with two levels. When analyzing the results, there are several possibilities: (1) There may or may not be a main effect for independent variable A, (2) there may or may not be a main effect for independent variable B, and (3) there may or may not be an interaction between the independent variables.
Page 216

[Figure 11.4](#) illustrates eight possible outcomes for a 2×2 factorial design. This figure might look a bit alarming at first, but we will walk you through its

interpretation. For each outcome, the means are given in a table and then graphed using a line graph. The means that are given in the figure are idealized examples: Such perfect outcomes rarely occur in actual research. Nevertheless, you should explore the graphs to determine for yourself why, in each case, there is or is not a main effect for A, a main effect for B, or an $A \times B$ interaction. Before you begin studying the graphs, it will help to think of concrete variables to represent the two independent variables and the dependent variable. You can think of anything you wish to represent the variables. Suppose you decide to use the example of crowding and group size. In this case, independent variable A could be crowding (A_1 is low crowding, operationalized as chairs placed 1 metre apart; A_2 is high crowding, chairs placed 1 centimetre apart), and independent variable B could be group size (B_1 is ten people per room; B_2 is two people per room). The dependent variable would be performance on a cognitive task, with higher numbers indicating better performance.

Figure 11.4 Outcomes of a factorial design with two independent variables



How can you tell what effects are present, given a graph? An interaction effect is easiest to spot: Are the two lines in the graph parallel to each other? If the lines are *not parallel*, then there *is* an interaction: The effect of one variable is different at different levels of the other variable. To detect main effects, it is easiest to examine the table rather than the graph, because the marginal means will stand out in a table. Examine whether each graph in [Figure 11.4](#) is depicting a main effect of factor A and/or a main effect of factor B.

The first four graphs in [Figure 11.4](#) illustrate outcomes in which there is no $A \times B$ interaction (note the parallel lines), and the last four graphs depict outcomes in which there is an interaction between the two independent variables (note the non-parallel lines). When there is an interaction, you need to carefully examine the means, which are given in the corresponding tables, to understand the specific pattern of the interaction. In some cases, there is a strong relationship between the first independent variable and the dependent variable at one level of the second independent variable, but there is no relationship (or only a weak relationship) at the other level of the second independent variable (e.g., panel 5). In other studies, the interaction may indicate that an independent variable has opposite effects on the dependent variable, depending on the level of the second independent variable. This pattern is shown in panel 8, the last graph in [Figure 11.4](#).

Test Yourself!

The independent and dependent variables in [Figure 11.4](#) do not have concrete variable labels. As an exercise, interpret each of the graphs using actual variables from two different hypothetical experiments (described below). This test works best if you draw the graphs on paper, including labels for the variables, separately for each experiment. For each hypothetical experiment, after you have labelled each graph with the variable names, consider whether the graph depicts a main effect of A, a main effect of B, and/or an $A \times B$ interaction. Once you have identified what effects are represented, try explaining in words what each graph would tell us about the relationships of those independent variables with the dependent variable.

You can try depicting the data as either line graphs or bar graphs. In general, line graphs are used when the levels of the independent variable on the horizontal axis (independent variable A) are quantitative and continuous, with low and high amounts. Bar graphs are more likely to be used when the levels of the independent variable are nominal, representing different categories (e.g., one type of therapy compared with another type).

Hypothetical experiment 1: Effect of age of defendant, and type of substance use during a crime, on sentencing. Participants read a scenario of a male, age 20 or 50, who was found guilty of causing a traffic accident while under the influence of either alcohol or marijuana, and then assigned him a jail sentence.

- Independent variable A: Type of offence (alcohol versus marijuana)
- Independent variable B: Age of defendant (20 versus 50 years of age)
- Dependent variable: Months of sentence (range from 0 to 10 months)

Hypothetical experiment 2: Effect of previous mood on the recall of humorous advertisements. Participants viewed a video that was either happy or sad, intended to induce a positive or negative mood. Next, they read print ads for eight different products over the next three minutes. These ads were either humorous or totally serious. The dependent variable was the number of ads correctly recalled.

- Independent variable A: Mood induction (happy versus sad video)
- Independent variable B: Ad content (humorous versus serious)
- Dependent variable: Number of ads recalled (range from 0 to 8)

Page 218



LO4 Breaking Down Interactions into Simple Main Effects

If you take a look at [Table 11.1](#) and [Figure 11.3](#) once again, there appears to be an interaction between the two independent variables. Whenever there is a significant interaction, we need to break it down further to understand it. The next step is to look at the simple main effects. A *simple main effect* is the mean difference at each level of one independent variable. This should not be confused with a main effect. The main effect of an independent variable takes the average *across* the levels of the other independent variable. With simple main effects, in contrast, the results are analyzed *within* each level of the other independent variable. Statistical tests would typically be used to identify whether simple main effects are statistically significant or not ([Chapter 13](#)).

Simple Main Effect of Independent Variable A

In [Figure 11.3](#) (or [Table 11.1](#), whichever you prefer), we can look at the simple main effect of A (food selection) *within* each level of B. Here, we compare the average food selections when the companion is either obese or thin. In this case, statistical tests showed that the simple main effect of food selection is not statistically significant when the eating companion is obese (means of 6.25 versus 4.26), but the simple main effect of food selection is statistically significant when the eating companion is thin (means of 9.82 versus 3.20) (McFerran et al., 2010). We ignore the marginal means of food selection, and instead interpret the cell means within each level of confederate body type.

Simple Main Effect of Independent Variable B

Alternatively, we could examine the simple main effect of B, body type, *within* each level of A. This will tell us whether the difference between the thin and obese eating companion is statistically significant when she eats 2 candies or 30 candies. In this case, the *simple main effect* of body type is not statistically significant when the eating companion eats 2 candies (means of 3.20 versus 4.26), nor is the *simple main effect* statistically significant when the eating companion eats 30 candies (means of 9.82 and 6.25).

Because these analyses overlap, you must choose to analyze only one of the simple main effects above (A or B, but not both). Which analysis you will be most interested in will depend on the predictions you made when you designed the study. The key point to remember here is that a significant interaction in a factorial design must be decomposed by examining cell means using a simple main effect analysis.

Variations on 2×2 Factorial Designs



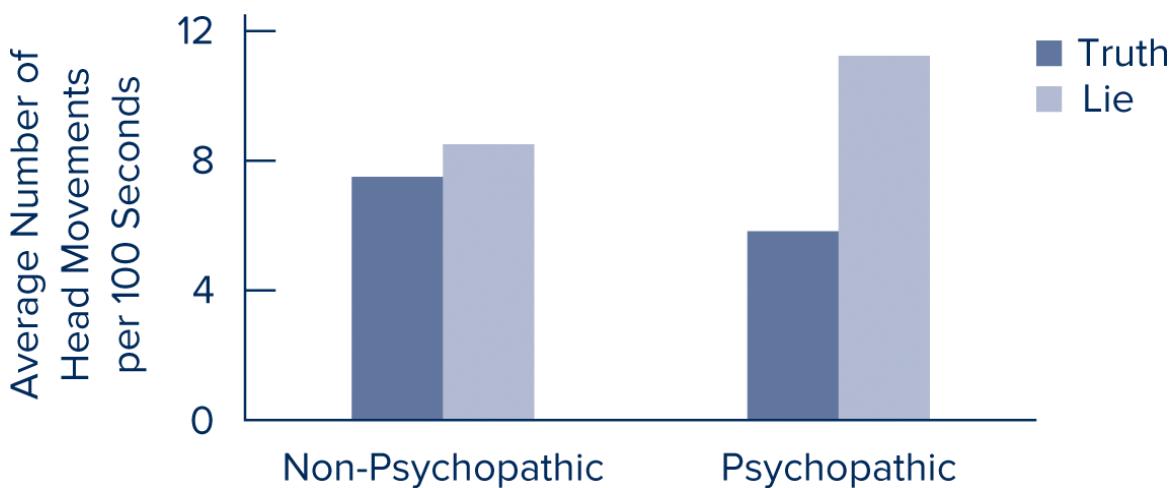
LO5 Factorial Designs with Manipulated and Non-manipulated Variables

One common type of factorial design includes both experimental (manipulated) and non-experimental (measured or non-manipulated) variables. This kind of design—sometimes called an independent variable by participant variable design, or IV × PV design—allows researchers to investigate how different types of people respond to the same manipulated variable. By different types of people, we mean individuals differing with respect to personal attributes such as age, ethnic group, personality characteristics, or clinical diagnostic category. These are known as participant variables. Importantly, participant variables cannot be randomly assigned or controlled, since participants bring these characteristics with them to the study. Just think how hard it would be to manipulate someone's chronological age! Therefore, the IV × PV design is not a true experiment, since it contains a variable that is measured and not manipulated. Page 219

The simplest IV × PV design includes one manipulated independent variable that has two levels and one participant variable with two levels. The two levels of the participant variable might be different age groups, groups that are low and high on a personality trait, or groups of short and tall individuals. An example of this design is a study investigating non-verbal behaviour among Canadian prison inmates (Klaver, Lee, & Hart, 2007). The psychopathy diagnosis of each inmate was the participant variable (either psychopathic or non-psychopathic), and the

story they were asked to tell was the manipulated independent variable (either a truth or a lie). One of the dependent variables was the number of head movements participants made while telling the story. The results are shown in [Figure 11.5](#). There was an interaction between psychopathy diagnosis and type of story on head movements observed. Overall, participants moved their heads more when telling a lie compared to telling the truth: In other words, there was a main effect of story condition. However, there was also an interaction between story and psychopathy diagnosis that reveals a more complex picture.

Figure 11.5 Interaction in IV × PV design



Adapted from Klaver, J.R., Lee, Z., & Hart, S.D. (2007). Psychopathy and nonverbal indicators of deception in offenders. *Law and Human Behavior*, 31, Figure 1, p. 345.



Among non-psychopathic participants, the number of head movements made was almost the same regardless of which story they were telling. However, psychopathic participants moved their heads much more often while telling a lie than when telling a truthful story. So head movements may not be very helpful when trying to tell whether a criminal is lying, unless that person is a psychopath!

Factorial designs that combine manipulated independent variables and participant variables are useful for investigating many interesting research questions. Fully understanding behaviour requires knowledge of both the situational variables that are able to be manipulated and the personal attributes of individuals. However, because participant variables can never be manipulated and randomly assigned,

we can never make causal claims when interpreting the results associated with participant variables.



LO6 Assignment Procedures and Sample Size

Techniques for assigning participants to conditions can be generalized to factorial designs. Recall from [Chapter 8](#) that in between-subjects designs different participants are randomly assigned to each of the conditions, whereas in a within-subjects design the same people participate in all conditions. These two types of assignment procedures have implications for the number of participants necessary to complete the experiment. We can illustrate this fact by looking at a 2×2 factorial design. The design can be completely between-subjects, completely within-subjects, or a [*mixed factorial design*](#), which is a combination of the two.

Page 220

Between-Subjects

In a 2×2 factorial design, there are four conditions. If we want a completely between-subjects (i.e., independent groups) design, different participants will be assigned to each of the four conditions. The food consumption study we described above is an example of this, with different people randomly assigned to each of the conditions (McFerran et al., 2010). For a 2×2 factorial design that is completely between-subjects, with 10 participants in each condition, we will need 40 different participants ([Figure 11.6](#)).

Figure 11.6 Number of participants (P) required for 10 observations in each condition in a 2×2 design

The figure consists of three separate 2x2 tables, each representing a different experimental design for a 2×2 factorial study. Each table has two columns labeled 'A' and 'B' at the top, and two rows labeled '1' and '2' on the left.

- I Between-Subjects Design:** This table shows two distinct groups of participants. Group A (rows 1, 3) contains participants P₁ through P₅. Group B (rows 2, 4) contains participants P₆ through P₁₀. The first column (A1) contains P₁, P₂, P₃, P₄, P₅ and P₂₁, P₂₂, P₂₃, P₂₄, P₂₅. The second column (A2) contains P₆, P₇, P₈, P₉, P₁₀ and P₂₆, P₂₇, P₂₈, P₂₉, P₃₀. The first row (B1) contains P₁₁, P₁₂, P₁₃, P₁₄, P₁₅ and P₃₁, P₃₂, P₃₃, P₃₄, P₃₅. The second row (B2) contains P₁₆, P₁₇, P₁₈, P₁₉, P₂₀ and P₃₆, P₃₇, P₃₈, P₃₉, P₄₀.
- II Within-Subjects Design:** This table shows a single group of participants (A1, A2) assigned to both conditions. Participants P₁ through P₅ are in both A1 and B1; participants P₆ through P₁₀ are in both A2 and B2.
- III Combination of Between-Subjects and Within-Subjects Designs:** This table shows a combination of designs. It has two groups (A1, A2). Group A1 contains participants P₁ through P₅. Group A2 contains participants P₁₁ through P₁₅. The first column (A1) contains P₁, P₂, P₃, P₄, P₅ and P₁₁, P₁₂, P₁₃, P₁₄, P₁₅. The second column (A2) contains P₆, P₇, P₈, P₉, P₁₀ and P₁₆, P₁₇, P₁₈, P₁₉, P₂₀. The first row (B1) contains P₁, P₂, P₃, P₄, P₅ and P₁₁, P₁₂, P₁₃, P₁₄, P₁₅. The second row (B2) contains P₆, P₇, P₈, P₉, P₁₀ and P₁₆, P₁₇, P₁₈, P₁₉, P₂₀.



Within-Subjects

In a completely within-subjects design (i.e., repeated measures), the same people will participate in all conditions. Suppose you have planned a study on semantic priming similar to the one by conducted by a group of University of Waterloo researchers (Ferguson, Robidoux, & Besner, 2009). Semantic priming happens when exposure to one concept (e.g., bird) activates a meaningfully related concept (e.g., wings). Let's consider part of the method that Ferguson and colleagues used to study this phenomenon. Participants read out loud words that were displayed, one at a time, on a computer screen, as quickly as possible. One independent variable was meaning: Words that appeared in sequence were either meaningfully related to each other (e.g., wood, tree) or not (e.g., tree, bath). Another independent variable was visual clarity: Words were presented in either a clear font or a fuzzy font. Both factors exhibited main effects on reading speed. People read faster when the words were related to each other compared to when they were not. Also, people read faster when words were clear as opposed to fuzzy. There was also evidence of an interaction: People read especially quickly when the words were related and clear, and read especially slowly when the words were unrelated and fuzzy.

In this completely within-subjects design, each person participated in all of the conditions by reading words in both clear and fuzzy fonts under both meaning conditions. If you wanted 10 participants in each condition, a total of 10 participants would be needed, as illustrated in the second panel in [Figure 11.6](#). Compared to a between-subjects design, this design offers considerable efficiency in terms of the number of participants required. Revisit [Chapter 8](#) to review other benefits as well as special challenges with within-subjects designs (e.g., order effects).Page 221

Mixed Factorial Design Using Combined Assignment

Earlier, we described a study on lie detection from head movements among psychopaths (Klaver et al., 2007), which illustrates the use of a mixed factorial design, with both between-subjects and within-subjects factors. The participant variable, psychopathy, is necessarily a between-subjects variable as participants could not be both psychopathic and non-psychopathic. The second independent variable, story truth, is a within-subjects variable: all participants told a truthful story and a lie.

The third panel in [Figure 11.6](#) shows the number of participants needed to have 10 per condition in a 2×2 mixed factorial design. In this table, independent variable A is a between-subjects variable (e.g., psychopathy diagnosis). Ten participants are assigned to level 1 of this independent variable, and another 10 participants are assigned to level 2. Independent variable B is within-subjects (e.g., story truth). The 10 participants assigned to A_1 receive both levels of independent variable B. Similarly, the other 10 participants assigned to A_2 receive both levels of the B variable. Thus, a total of 20 participants are required.

Increasing the Complexity of Factorial Designs

The 2×2 is the simplest factorial design. Building on this basic design, the researcher can arrange experiments that are more and more complex. One way to increase complexity is to increase the number of levels of one or more of the independent variables. Another way is to increase the number of independent variables. To describe these more complex designs, we will need to expand the general format for describing factorial designs, adding the additional variables:

- Number of levels of first IV \times Number of levels of second IV \times Number of levels of third IV . . . and so forth

A 2×3 design, for example, contains two independent variables: Independent variable A has two levels, and independent variable B has three levels. Thus, the 2×3 design has six conditions ($2 \times 3 = 6$). Following this logic, a 3×3 design has nine conditions. A $2 \times 2 \times 2$ design contains three independent variables, each with two levels, for a total of eight conditions ($2 \times 2 \times 2 = 8$).

Beyond Two Levels per Independent Variable

[Table 11.2](#) shows a hypothetical 2×3 factorial design with the independent variables of task difficulty (easy, hard) and anxiety level (low, moderate, high). The dependent variable is performance on the task. The numbers in each of the six cells of the design indicate the mean performance score across participants assigned to that group. Just like when we considered the 2×2 design, the marginal means (in the rightmost column and bottom row) show the main effects of each of the independent variables. The results in [Table 11.2](#) indicate a main effect of task difficulty because the overall performance mean in the easy-task group (i.e., 7.0) is higher than that for the hard-task group (i.e., 4.0). However, there is no main effect of anxiety because the mean performance score is the same in each of the three anxiety groups (i.e., 5.5). Is there an interaction between task difficulty and anxiety? Notice that increasing the amount of anxiety increases performance when the task is easy (means jump from 4 to 7 to 10), but decreases performance when the task is hard (means jump from 7 to 4 to 1). The

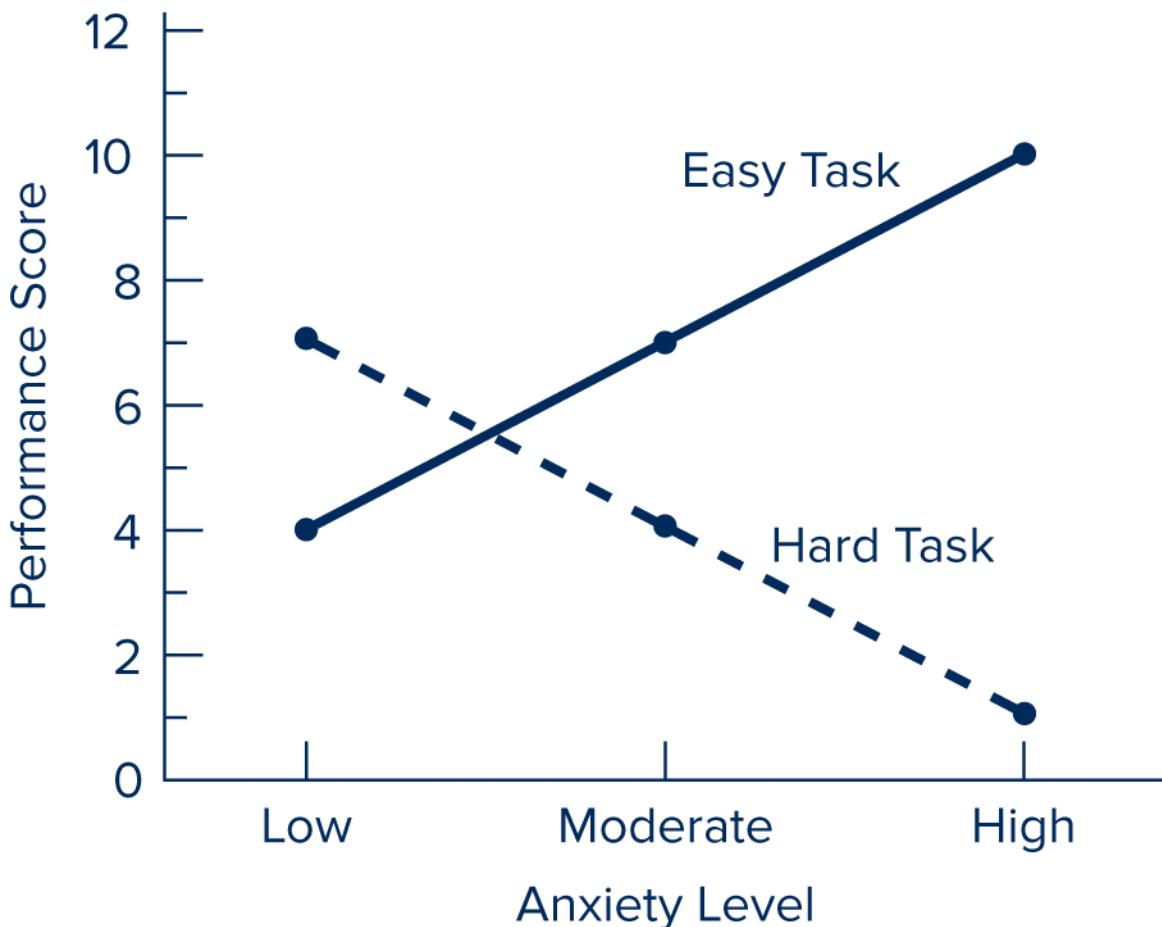
effect of anxiety is different, depending on whether the task is easy or hard; thus, there is an interaction.^{Page 222}

Table 11.2 2×3 factorial design

Task Difficulty	Anxiety Level			Marginal Means (shows main effect of task difficulty)
	Low	Moderate	High	
Easy	4	7	10	7.0
Hard	7	4	1	4.0
Marginal Means (shows main effect of anxiety level)	5.5	5.5	5.5	

This interaction can also be presented as a graph. [Figure 11.7](#) is a line graph in which the solid line shows the effect of anxiety for the easy task and the dotted line represents the effect of anxiety for the hard task. As noted previously, line graphs are used when the independent variable represented on the horizontal axis is quantitative (i.e., the levels of the independent variable are increasing amounts of the variable).

Figure 11.7 Line graph of data from 2 (task difficulty) \times 3 (anxiety level) factorial design



Beyond Two Independent Variables

We can also increase the number of variables in the design. A $2 \times 2 \times 2$ factorial design contains three variables, each with two levels. Thus, there are eight conditions in this design. In a $2 \times 2 \times 3$ design, there are 12 conditions, and in a $2 \times 2 \times 2 \times 2$ design, there are 16. The rule for constructing factorial designs remains the same throughout.

An example $2 \times 2 \times 2$ factorial design is constructed in [Table 11.3](#). The independent variables are (1) instruction method (online, face-to-face), (2) class size (small, moderate), and (3) year of university (first, fourth). Note that year of university is a participant variable and the other two variables are manipulated variables. The dependent variable is performance on a standard test.

Table 11.3 $2 \times 2 \times 2$ factorial design

		<i>Class Size</i>
		<i>Instruction method</i> Small (30) Moderate (80)
<i>First Year Students</i>		
Online		
Face-to-face		
<i>Fourth Year Students</i>		
Online		
Face-to-face		

Notice that the $2 \times 2 \times 2$ design can be seen as two 2×2 designs, one for the first year students and another for the fourth year students. The design yields three possible main effects: one for each independent variable. For example, let's consider the possible main effect of class size, which would be examined in the bottom row margin (not depicted). The overall mean for the small class size is obtained by considering all participants who experience the small class, regardless of instruction method or year of university. Similarly, the moderate class size mean is derived from all participants in this condition. The two marginal means (small versus moderate class size) are then compared to see whether there is a significant main effect: Is one class size superior to the other, overall?Page 223

The design also allows us to examine interactions. In the $2 \times 2 \times 2$ design, there could be interactions between (1) method and class size, (2) method and year of university, and/or (3) class size and year of university. There could also be a three-way ("higher-order") interaction that involves all three independent variables. Here, we want to determine whether the nature of the interaction between two of the variables differs depending on the particular level of the other variable. Three-way interactions are complicated and are less common in behavioural science than are two-way interactions.

Sometimes, new researchers are tempted to include many independent variables in a single study. This practice is problematic because the design may become needlessly complex and require enormous numbers of participants, especially with between-subjects designs. The $2 \times 2 \times 2$ design had 8 groups; a $2 \times 2 \times 2 \times 2$ design has 16 groups; adding yet another independent variable with two levels means that 32 groups would be required. Also, when there are more than two independent variables, analyses become increasingly complex, and some of the

particular conditions that are produced by the combination of so many variables may not make sense.

Experimental designs all use the same logic for determining whether the independent variable(s) cause a change on the dependent variable. In the next chapter, we will consider alternative designs that use somewhat different procedures for examining relationships between variables under special circumstances (e.g., when random assignment to condition is impossible).



Illustrative Article: Complex Experimental Designs

We have all seen people walking while using a mobile device. Often they are talking, but they might be texting or viewing something on a smartphone screen. The authors of this illustrative article were interested in the “minieounters” that occur when two people pass each other on a walkway. How does mobile device use affect what happens in these encounters?

Patterson, Lammers, and Tubbs (2014) conducted an experiment in which confederates varied their greeting of a stranger approaching with or without a mobile device.

First, acquire and read the article:

- Patterson, M. L., Lammers, V. M., & Tubbs, M. E. (2014). Busy signal: Effects of mobile device usage on pedestrian encounters. *Journal of Nonverbal Behavior*, 38, 313–324. doi:10.1007/s10919-014-0182-4

Page 224 Then, after reading the article:

1. Identify the design of the experiment.
2. Identify each independent variable. For each independent variable:
 1. What were the levels (values) of the variable?
 2. Was the variable a manipulated variable or a participant variable?
 3. How was the variable operationalized?

3. Identify each dependent variable. Describe how each dependent variable was operationalized.
4. Which dependent variable was discarded from the analysis? Why?
5. Describe the results in Figure 2. Is there an interaction effect?
6. Describe at least one additional result that you found interesting or surprising.

Study Terms

Test yourself! Define and generate an example of each of these key terms.

- cell (p. 213).
- factorial design (p. 212)
- interaction (p. 213)
- IV × PV design (p. 218).
- levels (p. 210)
- main effect (p. 213)
- marginal mean (p. 213)
- mixed factorial design (p. 219)
- moderator variable (p. 215)
- simple main effect (p. 218)

Review Questions

Test yourself on this chapter's learning objectives. Can you answer each of these questions?

1. Why would a researcher use more than two levels of the independent variable in an experiment?
2. What is a factorial design? Why would a researcher use a factorial design?
3. What are main effects? What is an interaction?
4. How do you use a graph to estimate whether there are any main effects, or if there is an interaction?
5. Cover the “yes” and “no” answers in [Figure 11.4](#). For each graph, identify whether there is a main effect of factor A, a main effect of factor B, and/or an interaction between factors A and B.
6. What information does a simple main effect analysis provide? How does it differ from a main effect?
7. Generate an example of an $IV \times PV$ factorial design. When might this design be useful?
8. What two pieces of information do you need to identify the number of conditions (or cells) in a factorial design?

Deepen Your Understanding

Develop your mastery of these concepts by considering these application questions. Compare your responses with those from other people in your study group.

1. In a study by Chandler and Schwarz (2009), participants read a description of a man named Donald. The description was ambiguous enough that his behaviour could be interpreted as being either assertive or hostile. Researchers were interested in whether a certain hand movement could prime the concept of hostility and influence personality ratings of Donald. Therefore, while reading about Donald, participants were asked to engage in a “motor task” that involved either extending their index finger or their middle finger upward (the latter is a signal of aggression in North American culture). Then participants rated Donald’s personality on two trait dimensions: aggressive traits (e.g., hostile, unfriendly) and unrelated, control traits (e.g., intelligent, boring).
 1. Identify the independent variable(s) and dependent variable(s).
 2. Identify the design of this experiment.
 3. How many conditions are in the experiment?
 4. Is there a within-subjects variable in this experiment? If so, what are the levels?
 5. Is there a participant variable? If so, identify it. If not, can you suggest a participant variable that might be included?
2. Chandler and Schwarz (2009) reported the following mean personality ratings (higher numbers indicate greater levels of the trait): Middle finger—Aggressive traits (8.41), Middle finger—Unrelated traits (6.61), Index finger—Aggressive traits (6.74), and Index finger—Unrelated traits (6.38). Assume there are equal numbers of participants in each condition.

1. Graph the means.
 2. Are there any main effects?
 3. Is there an interaction?
 4. Describe the results in a few sentences.
3. Assume that you want 15 participants in each condition of your experiment, which uses a 3×3 factorial design. How many different participants do you need for (a) a between-subjects design, (b) a within-subjects design, and (c) a mixed factorial design?
4. Read the following research scenarios and fill in the correct answer in each column of the table. It may be helpful to create a table or a graph like the ones depicted in this chapter.

Number of Independent Variables	Number of Experimental Conditions	of Possible Main Effects	Number of Possible Interactions
--	--	-----------------------------------	--

	Number of Independent Variables	Number of Experimental Conditions	of Possible Main Effects	Number of Possible Interactions Effects
1. Participants were randomly assigned to read a short story printed in either 12-point or 14-point font in one of three font styles: Calibri, Times New Roman, or Arial. Afterwards, they answered several questions designed to measure memory recall.				

	Number of Independent Variables	Number of Experimental Conditions	of Possible Main Effects	Number of Possible Interactions Effects
2. Researchers conducted an experiment to examine sex and physical attractiveness biases in juror behaviour. Participants were randomly assigned to read a scenario describing a crime committed by either an attractive or unattractive woman, or an attractive or unattractive man. The criminal was described as overweight or average weight.	2	2	2	2

Chapter 12

Page 226

Descriptive Statistics: Describing Variables and the Relations among Them



©Digital Images Studio/Shutterstock

How would you describe this gorgeous creature, in one sentence and with great accuracy? The first step of data analysis is getting a good sense of the data, describing it well through descriptive statistics.

Learning Objectives

Keep these learning objectives in mind as you read to help you identify the most critical information in this chapter.

By the end of this chapter, you should be able to:

1. [LO1](#) Describe frequency distributions and ways to visualize them.
2. [LO2](#) Interpret measures of central tendency and variability.
3. [LO3](#) Identify ways to describe and graph relationships involving nominal variables.
4. [LO4](#) Identify and interpret an estimate of effect-size comparing two groups.
5. [LO5](#) Interpret correlation coefficients and scatterplots to describe relationships among continuous variables.
6. [LO6](#) Describe how regression equations and multiple correlations are used to make predictions.
7. [LO7](#) Discuss how a partial correlation addresses the third-variable problem.

Page 227 Statistics help us understand the data we collect during research. In this chapter, we explore ways in which statistics and graphs are used to describe and represent our data. In [Chapter 13](#), we discuss how statistics are used to help us make inferences about the populations from which our samples are drawn. In these chapters we introduce the underlying logic and general procedures for conducting statistics, with the specific calculations for various statistics provided in [Appendix B](#).

Concepts throughout this chapter are frequently illustrated using data from the World Values Survey (World Values Survey Wave 6, 2010–2014). These data were collected from over 74,000 participants across 52 countries, and address a wide variety of topics, including life satisfaction; perceived health, wealth, and safety; and religious and political views. You can download the dataset at www.worldvaluessurvey.org and use it to practise what you learn here.

Revisiting Scales of Measurement

Choosing which statistical analyses and graphs are most appropriate depends on each variable's scale of measurement ([Chapter 5](#)). A brief review of the different scales of measurement (i.e., nominal, ordinal, interval, and ratio scales) will help us throughout the next two chapters.

The levels of a *nominal* scale are simply different categories or groups that have no intrinsic numerical properties. For example, two different kinds of therapies for depression would be nominal categories. Variables using an *ordinal* scale rank order the levels from lowest to highest (or least to most), but the intervals between each rank order are not equal. A list of the top ten restaurants in Halifax would use an ordinal scale. For an *interval* scale, the distances between each level are equivalent in size. In this case, the difference between 90 and 95 on the scale should be the same as the difference between 115 and 120. Scores on an intelligence test are an example of an interval scale. However, there is no meaningful zero point that indicates a total “absence” of the construct. In contrast, *ratio* scales have equal intervals in addition to a true zero point. Response time and age are examples of ratio scales.

In the behavioural sciences, it is sometimes difficult to know precisely whether an ordinal or an interval scale is being used. For example, we assume that asking people to rate their state of health on a 4-point scale uses an interval scale, when the points are labelled *very good*, *good*, *fair*, and *poor*. Yet it is difficult to claim that the difference between *very good* (4) and *good* (3) is the same as the difference between *fair* (2) and *poor* (1). However, it is common practice to treat variables measured like this as an interval scale, because when ordinal scales are averaged across many instances (e.g., many items in a self-report scale), they take on properties similar to an interval scale.

The statistical procedures used to analyze data with interval and ratio variables are identical. Importantly, data measured on interval and ratio scales can be summarized using an arithmetic average, or what is known as

the *mean*. It is possible to provide a number that reflects the mean amount for these variables. For example, “on average, people rated their health as 2.1 on a 4-point scale,” or the “mean age of the sample was 20.9 years.” As we will see, the mean is used in many statistical analyses. Variables measured on interval and ratio scales are often referred to as continuous variables because the values represent an underlying continuum (at least in theory). Because variables using interval and ratio scales can be treated the same way statistically, we will group them together and refer to them as continuous variables.

Describing Each Variable

After collecting data, it is time to analyze these data and view our results. The first step a researcher should take is to explore each variable separately. Doing so allows us to get a sense for what the data for each of our variables look like and also identify any possible errors that might have occurred during data collection (e.g., a 6 on a 7-point scale accidentally recorded as 66). What graphs and analyses we should use will depend on the measurement scale. Let's consider some options.



LO1 Graphing Frequency

Distributions

A *frequency distribution* indicates the number of participants who receive or select each possible score on a variable. Frequency distributions can be created for variables using any scale of measurement. You are likely familiar with frequency distributions if your professor has ever presented a graph showing how many students got each score on an exam.

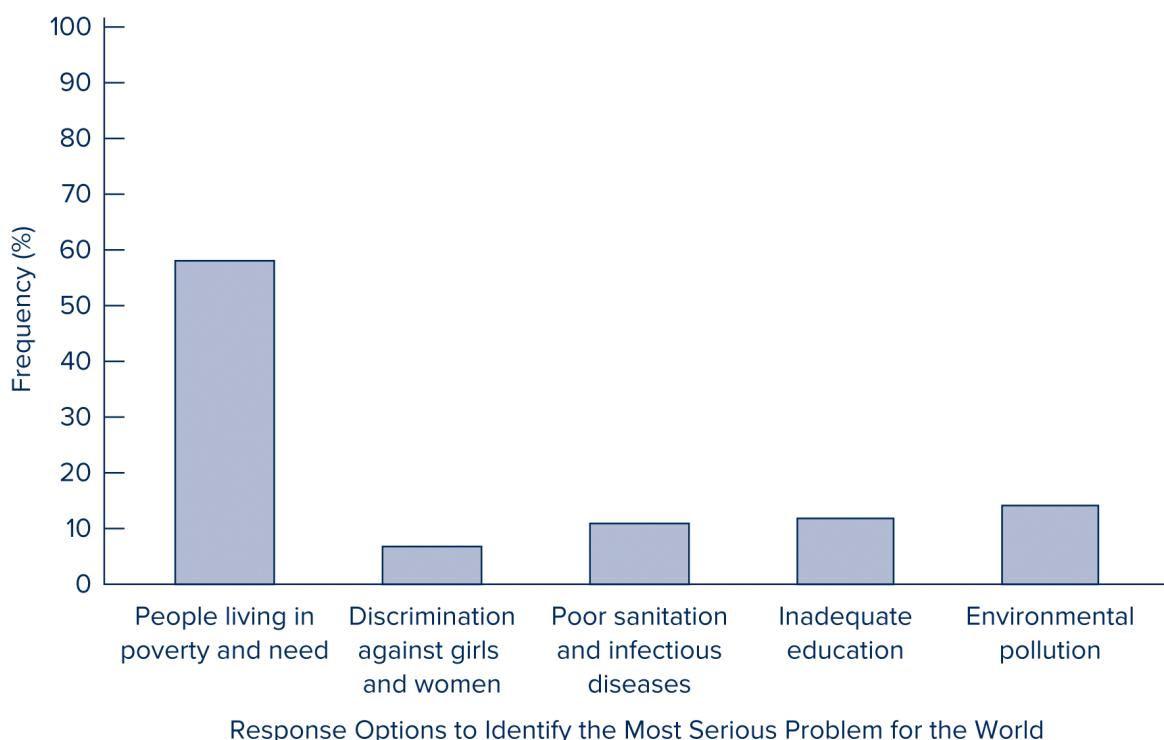
Graphical representations of frequency distributions allow us to see at a glance what our data look like. You can quickly see what scores are most common (or frequent), which are infrequent, and the shape of the distribution. You can also tell whether there are any *outliers*: scores that are unusual, unexpected,

impossible, or very different from the scores of other participants. An outlier might reflect a data-entry error that can be corrected (e.g., a person's age appearing as 333). Let's examine several types of graphs that are used to depict frequency distributions: the bar graph, pie chart, histogram, and frequency polygon. Of these types, bar graphs and histograms are the most common.

Bar Graphs

A *bar graph* uses a separate and distinct bar for each piece of information. Bar graphs are commonly used for comparing group means (e.g., [Figure 12.5](#)) but can also be used for comparing percentages, as we have shown in [Figure 12.1](#). In this figure, and all subsequent examples, we see data from respondents to the World Values Survey. These individuals each chose one of five responses to the question, "Which of the following problems do you consider the most serious one for the world as a whole?" In this graph, the horizontal *x*-axis shows the five possible responses, and the vertical *y*-axis shows the proportion of people who chose each response. Over half of the sample chose poverty over the other options.[Page 229](#)

Figure 12.1 Most serious problem for the world as a whole (World Values Survey)





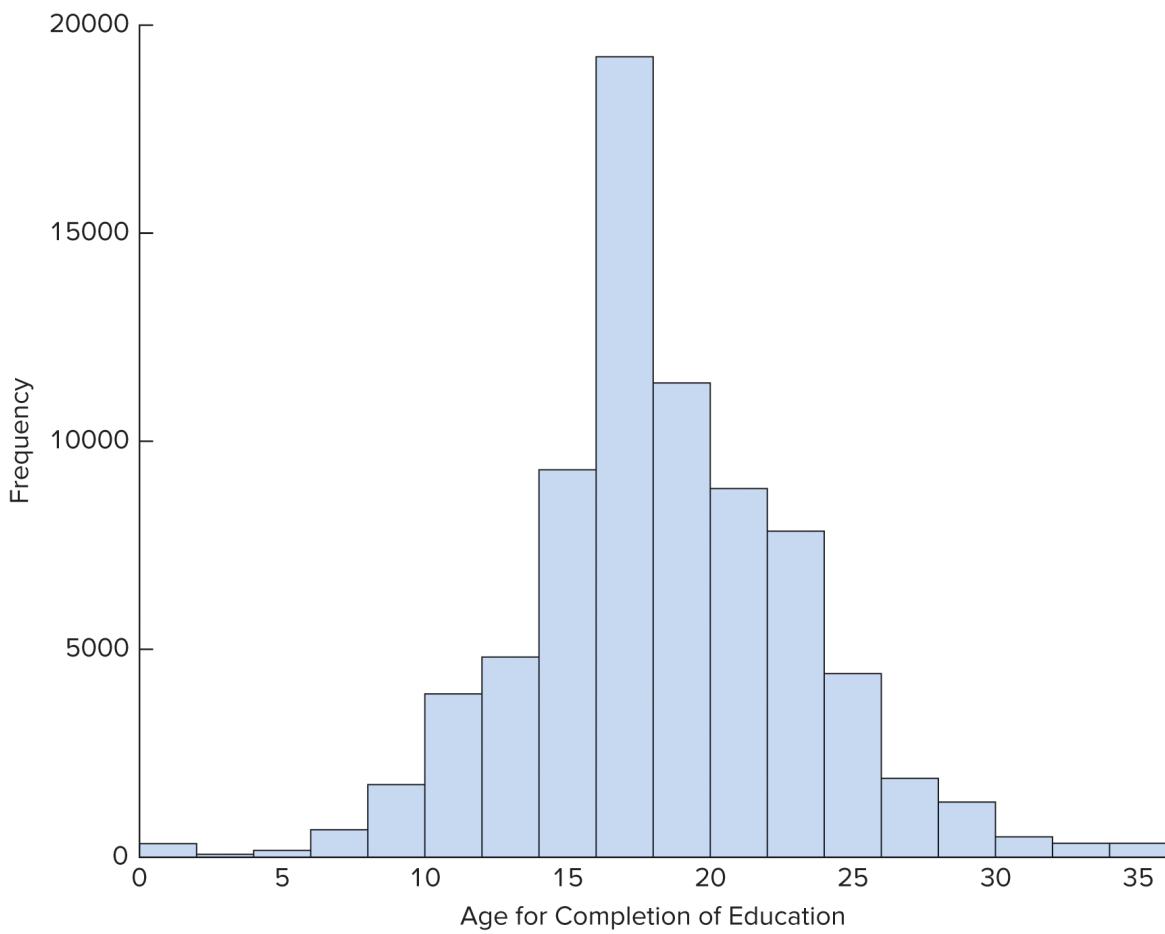
Pie Charts

A *pie chart* divides a whole circle, or “pie,” into “slices” that represent relative percentages. Pie charts are particularly useful when representing data on a nominal scale. Instead of using a bar graph for the data in [Figure 12.1](#), we could have divided a circle into five sections corresponding to the five response options, with these responses being nominal (or categorial) in nature. The largest section would represent the poverty response because it was selected most frequently. You will not see many pie charts in the journal articles that you read. However, they are used frequently in applied research reports, infographics, newspapers, and magazines.

Histograms

A *histogram* uses bars to display a frequency distribution for a continuous variable. In this case, the values along the x -axis are continuous and show increasing amounts of a variable, such as blood pressure, reaction time, or number of correct responses. [Figure 12.2](#) shows a histogram summarizing the age at which respondents completed their education. Notice that the histogram bars touch each other, reflecting the fact that the variable on the x -axis is a continuous variable. This display contrasts with the bar graph ([Figure 12.1](#)), which has gaps between each bar helping to communicate the fact that the values on the x -axis are nominal categories (i.e., not continuous).

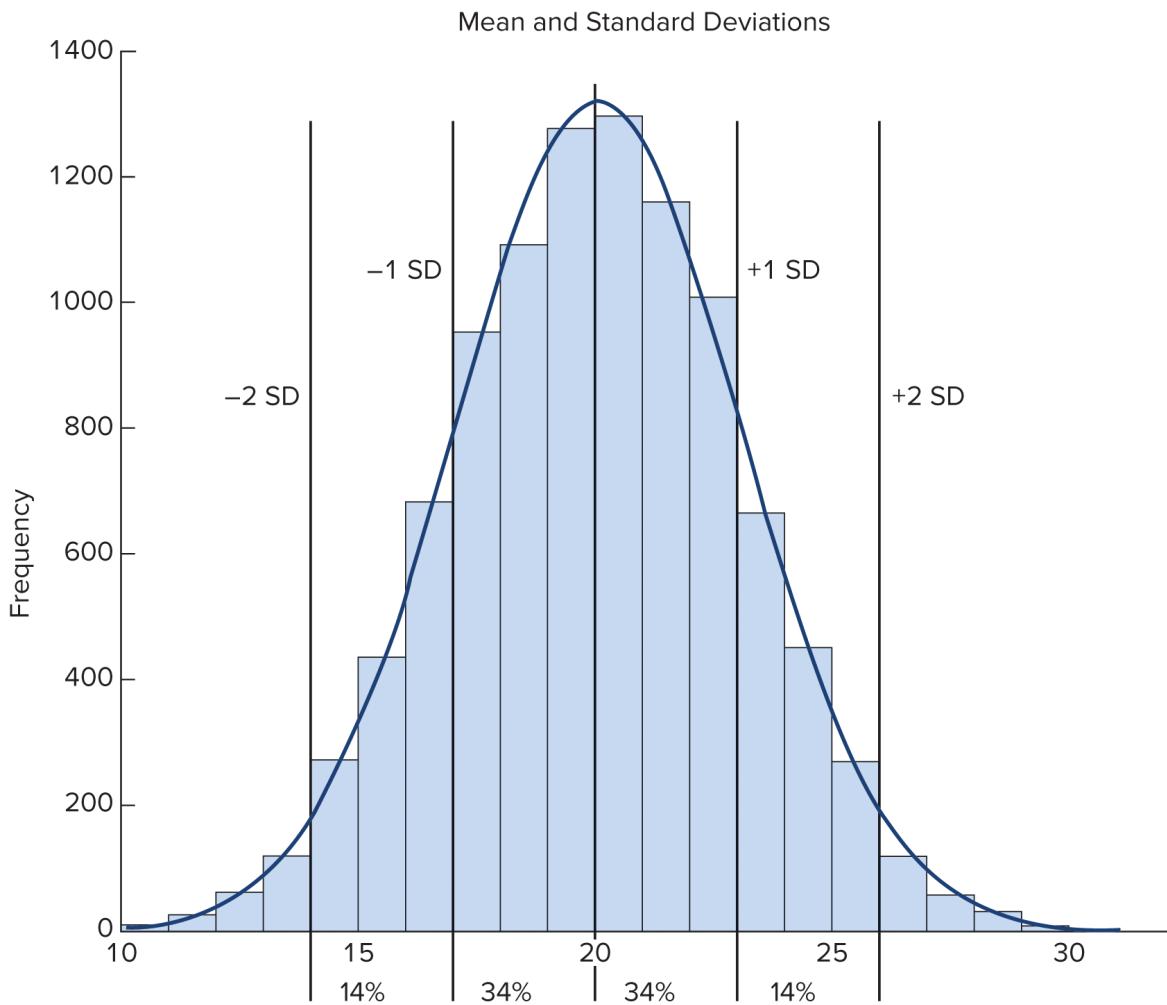
Figure 12.2 Histogram showing frequency of age at which respondents completed their education, from the World Values Survey



You might recognize the shape of this distribution, as a “bell-shaped curve.” This is known as a [normal distribution](#), a distribution of scores that is frequently observed, and rather important for statistics. In a normal distribution, the majority of the scores cluster around the [mean](#) (M , also known as the average, discussed in detail below), with fewer and fewer scores observed the further you get from the mean. This distribution is only possible for continuous variables (i.e., interval or ratio scales), and this distribution is frequently observed for many naturally occurring variables (e.g., height of dogs, weight of cats, length of ferrets). If you think about it, this makes intuitive sense as, for many things, most observations are around the average, and it is uncommon to observe examples far from the average. There are, for example, very few cats that weigh 18 kilograms (39 lbs.) or dogs that are over 1 metre tall (39 in.). The normal distribution is important because if our sample is drawn from a population of scores that are normally distributed, then we know a lot about this distribution. For example, we know how many scores fall within 1 or 2 (or 3) standard deviations from the mean. The

standard deviation (SD) is a common measure of variability, or how the scores are spread out with respect to the mean (i.e., how far each score is from the mean, on average; discussed further below). [Figure 12.3](#) shows some generated data, with the curve for an ideal normal distribution superimposed on top.[Page 230](#)

Figure 12.3 The ideal normal distribution, with percentage of data falling within 1 and 2 standard deviations of the mean



As you can see from this figure, the majority of scores are clustered around the mean of 20. In this example, the standard deviation is 3, and in a normal distribution, about 68 percent of scores (68.1 percent, actually) fall within 1 standard deviation above and below the mean (i.e., between 17 and 23). Roughly 96 percent of the data (95.4 percent, in truth, but who's counting?) fall within 2 standard deviations above and below the mean (i.e., between 14 and 26). So in an

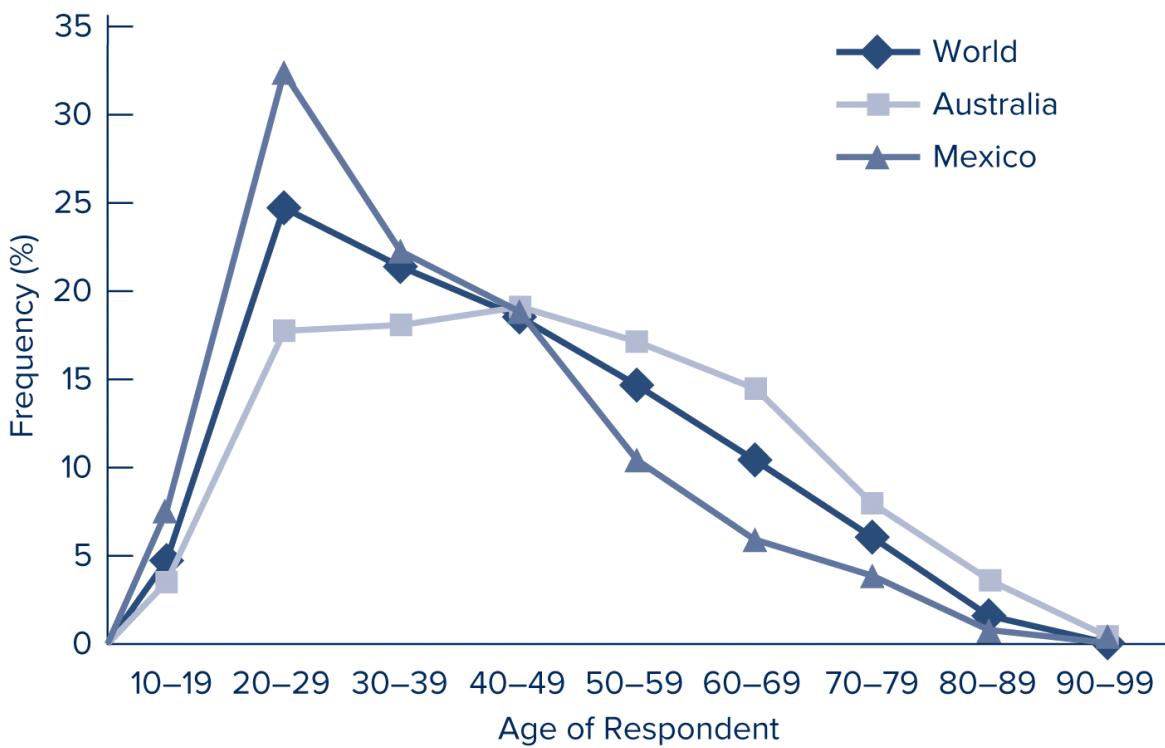
idealized normal distribution, a good chunk of the scores are within 1 *SD* of the mean (more than two-thirds), and almost all of the data is within 2 *SDs* of the mean. From this we gain an intuitive sense that few scores are greater than 1 *SD* from the mean, and even fewer appear more than 2 *SDs* from the mean.

When we use descriptive statistics such as a histogram to visualize our sample data, we often want to figure out if our sample is drawn from a population that is normally distributed, because this will determine what statistics we should use. Many of the statistics that we will learn about (e.g., the *t*-test, *F*-test) should only be used if our sample data are drawn from a normally distributed population. (These statistics are known as parametric statistics.) Thus, when looking at our data, we should first determine if this is the case, using descriptive statistics, because if our data are not from a normal distribution we will have to use a different set of statistics (i.e., non-parametric statistics).Page 231

Frequency Polygons

Frequency polygons, an alternative to histograms, use a line to represent frequencies for continuous variables (i.e., interval or ratio scales). Frequency polygons are especially helpful when you want to examine frequencies for multiple groups simultaneously. Age data are depicted in [Figure 12.4](#) using three frequency polygons: one for the total sample, one for the Australian sample, and one for the Mexican sample. This graph helps us to visualize the distribution of age in each group. When interpreting responses from different countries, it may be useful to note that the Australian sample represents an approximately even proportion of people across ages 20 through 59, but the Mexican sample includes more 20- to 29-year-olds than any other group. Layering frequency polygons in the same graph makes it easy to see this difference.

Figure 12.4 Frequency polygons illustrating the distributions of respondents by age for the full World Values Survey sample, for the Australian sample, and for the Mexican sample



Note: Each frequency polygon is anchored at scores that were not obtained by anyone (e.g., 0 years), just beyond the range of collected data.



LO2 Descriptive Statistics

Visualizing frequency distributions are a good way to take an initial look at our data in order to gain a sense of it, but we can also calculate statistics to describe or summarize our data. These statistics are known as *descriptive statistics*. Two main types of descriptive statistics are (1) measures of central tendency and (2) measures of variability. Measures of central tendency try to capture how participants scored overall, across the entire sample, in various ways. In contrast, measures of variability attempt to summarize how differently the scores are from each other, or how widely the scores are spread out or distributed. Both types of descriptive statistics help us to summarize the scores for a variable. If a study makes a comparison between groups (e.g., an experiment with two conditions, participants' cultural background), descriptive statistics are calculated separately for each group.

Central Tendency

A *central tendency* statistic tells us what the scores are like as a whole, or how people scored on average. There are three measures of central tendency: (1) the mean, (2) the median, and (3) the mode. Each of these statistics have different strengths and weaknesses, and also vary in how appropriate they are in different situations (e.g., the scale of measurement of the variable).Page 232

As noted above, the *mean* of a set of scores is obtained by adding all the scores together and then dividing this number by the number of scores. It is likely the measure of central tendency that you are most familiar with, and often what people mean when they talk about an “average.” For calculations, it is represented by the symbol X (pronounced “X bar”), and in scientific reports, it is abbreviated as M . The mean is only appropriate when analyzing scores that use an interval or ratio scale (i.e., continuous variables), because the actual values are used and so these values must be numerically meaningful. The mean age of all respondents to the World Values Survey was 41.9 years. For Australian respondents, the mean age was 46.4 years, and for Mexican respondents it was 37.5 years. The mean provides us with a single value that summarizes age for each group. We will continue to use these data from the World Values Survey as an example in discussing measures of central tendency and variability.

Another measure of central tendency is the *median*, which is the score that divides the group in half (with 50 percent scoring below and 50 percent scoring above the median). For an odd number of scores, it is easy to identify which score falls right in the middle. When there are even number of scores, the middle will fall between two different numbers, so we take the mean of these two values. In

scientific reports, the median is abbreviated as *Mdn*. The median can be calculated for continuous variables, just like the mean. In addition, it is also appropriate when scores are on an ordinal scale, because it takes into account only the rank order of the scores. The median age of all respondents to the World Values Survey was 40. The median for Australian respondents was 45, but it was only 35 for Mexican respondents.

The *mode* is the most frequent score, or the score that is most common. The mode can be calculated for variables that employ an interval, ratio, or ordinal scale. In addition, it is the only measure of central tendency appropriate for scores on a nominal scale. For example, in [Figure 12.1](#), note that the mode for this question about the most serious problem facing the world was “poverty.” We would therefore refer to this as the modal response. The mode does capitalize on most of the actual scores collected, but simply indicates the value that occurs most often (or values, if there is a tie for most frequent). The most frequently occurring age among respondents for these data is 30 years (44 among Australians, 29 among Mexicans). If the data are perfectly normally distributed, the mean, the median, and the mode will be equal (i.e., have the same value). Think about the normal distribution, look at [Figure 12.3](#), and it should make sense why the average, the middle, and the most common score will be identical for normally distributed data. However, if the data deviate from a normal distribution, then you will often find that these different measures of central tendency will be different.

So which measure of central tendency should you use: the mean, median, or mode? Each gives you different information and so answers a slightly different question; the measure you will focus on will therefore depend on what you are interested in knowing. Do you want to know which choice was most popular in your sample? In this case you should use the mode. Or if you want to incorporate all of your data and understand what the average response was, then the mean would be more appropriate. As a general guideline, it is good to calculate all of the measures of central tendency that are appropriate for the variable (based on its scale of measurement). Seeing these results and thinking about whether they differ and perhaps why is a great way to engage with your data. One additional note of caution: The mean is very susceptible to outliers. If there are scores that are very different from most other scores, either very high or very low, then the mean can be misleading. In this case, the median would be the more appropriate measure of central tendency. For example, the median family income of a country is usually a better measure of central tendency than the mean family income. Because a relatively small number of people have extremely high incomes, using the mean would make it appear that the “average” person makes more money

than is actually the case. Lastly, when deciding which measure of central tendency to use, we need to consider whether the data are normally distributed or not. If the data are not normally distributed, not symmetrical with most data near the mean, but instead the distribution is skewed to one side or another, then the mean will also be skewed. This makes the median a better representation of central tendency.

Page 233

Table 12.1



TEST YOURSELF! Calculate the mean, median, and mode for each set of data (answers at bottom of the page)

Data	Mean	Median	Mode
1, 3, 2, 9, 9	►	►	►
	Answer	Answer	Answer
10, 7, 8, 7, 6, 892	►	►	►
	Answer	Answer	Answer
2, 1, 2, 8, 2, 10, 2	►	►	►
	Answer	Answer	Answer
MJ, Mike, Faraz, Jorge, Carlos, Dan, Sana, TJ, Mike, Estefan, Jodi, Amogh, Rob, Shahrin, Annie, Alex, Sandeep, James, Petra	►	►	►
	Answer	Answer	Answer

Variability

A measure of *variability* characterizes the amount of spread in a distribution of scores, for continuous variables (i.e., interval or ratio scales). One common measure of variability is the standard deviation, which indicates how far away the scores tend to be from the mean, on average. In scientific reports, it is abbreviated as *SD*, and in formulas, it is abbreviated as *s*. The standard deviation is small when most people have scores close to the mean, and it becomes larger as more people have scores further from the mean. For example, among Australian respondents, the standard deviation of age is 17.7. Among Mexicans, the standard deviation of age is 15.2, meaning that the distribution of these ages is clustered more tightly around the mean relative to the Australian sample, to a small degree. Together, the mean and standard deviation provide information about the way the scores are distributed.

The standard deviation is derived by first calculating the *variance*, symbolized as s^2 (the standard deviation is the square root of the variance; go to [Appendix B](#) for equations). Note that, as with the mean, the calculation of the standard deviation (and variance) uses the actual values of all scores. As a result, the standard deviation is only appropriate for continuous variables (i.e., interval and ratio scale variables). Reading closely the formula for calculating the standard deviation (and the variance), and practising some of these calculations, will give you a strong understanding of what is meant by variability and what these measures of variability represent ([Appendix B](#)). It is strongly recommended that you calculate the standard deviation for some data by hand, at least once, in order to gain a proper grasp of these measures of variability.

Another measure of variability is the range. In its most commonly used calculation, the range is the difference between the highest score (i.e., the maximum; *Max.*) and the lowest score (i.e., the minimum; *Min.*). The age range for Australian respondents is 77 (*Max.* = 95, *Min.* = 18), whereas it is 75 among Mexican respondents (*Max.* = 93, *Min.* = 18).



LO3 Describing Relationships Involving Nominal Variables

A great deal of research focuses on the study of relationships between variables. Depending on the way the variables are measured, and the research questions being explored, there are different ways to describe how variables are related. In this section, we will consider situations that involve nominal variables (i.e., categorical variables). Here we explore two approaches, one that involves comparing group percentages (when both variables are nominal) and one comparing group means (when one variable is nominal and the other is continuous). After describing data using percentages or means, the next step is typically a statistical analysis to determine whether there is a statistically significant difference between nominal groups ([Chapter 13](#) and [Appendix B](#)).

Comparing Groups of Participants

Comparing Group Percentages

Suppose you want to know whether Australians and Mexicans differ in what they consider to be the most serious problem facing the world. In the World Values

Survey, people from each country were asked to choose one of five options ([Figure 12.1](#)). How many people chose “discrimination against girls and women” as the most serious problem? Sixty-two of the 1,472 Australian respondents chose this option, whereas 283 of the 1,987 Mexican respondents did. To describe these results, we can calculate the percentage for each group: 4.2 percent of Australian respondents versus 14.2 percent of Mexican respondents. There appears to be a relationship between nationality and perceived seriousness of discrimination against girls and women (at least when comparing these two nationalities, specifically). Note that we are focusing on percentages because both variables (nationality and choice of most serious problem) are nominal.

Comparing Group Means

Many studies are designed to compare mean responses to continuous variables made by participants in two or more nominal groups (including the designs found in [Chapters 8, 10](#), and some in [Chapter 11](#)). For example, consider an experiment designed to study the effect of feeling powerful on feelings of well-being (Kifer, Heller, Perunovic, & Galinsky, 2013). All participants wrote about a situation from their past, but participants were also randomly assigned to one of two conditions. Those in the high-power condition wrote about a time when they were able to control or evaluate another person. In contrast, those in the low-power condition wrote about a time when someone else controlled or evaluated them. Each participant then rated their well-being. Researchers found that the mean well-being scores reported by people in the high-power condition were higher than for those in the low-power condition. Looking to improve your well-being today? Consider remembering a time when you felt powerful!

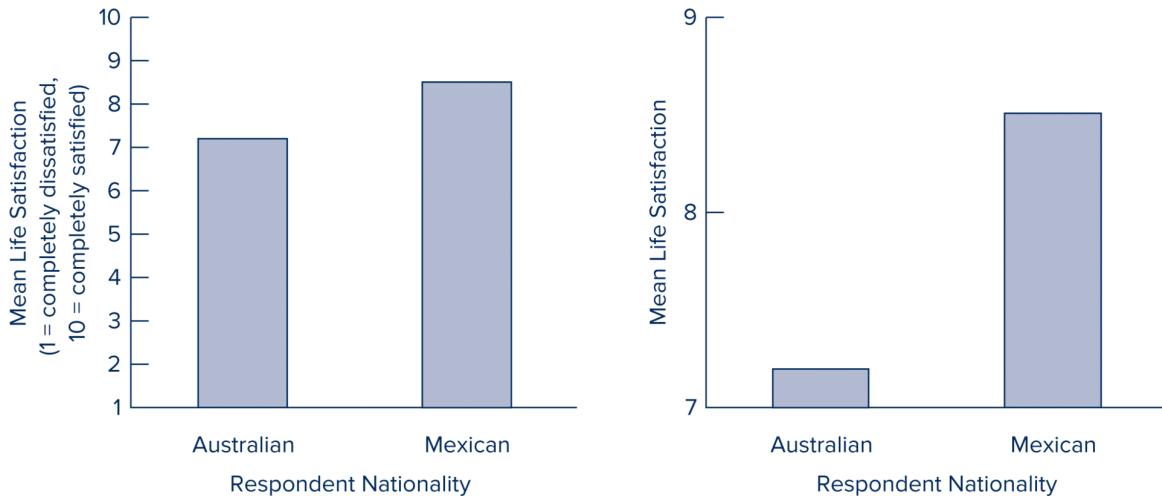
Another way to form groups for comparison is using participant variables ([Chapter 11](#)). Using data from the World Values Survey, we can compare the mean life satisfaction for different groups. One question asks, “All things considered, how satisfied are you with your life as a whole these days?” Response options ranged from 1 (*completely dissatisfied*) to 10 (*completely satisfied*). The mean response for Australian respondents was 7.2; for Mexican respondents, it was 8.5.[Page 235](#)

Graphing Nominal Data

A common way to graph relationships between variables when one variable is nominal is to use a bar graph or a line graph ([Chapter 11](#)). [Figure 12.5](#) is a bar

graph depicting the mean life satisfaction scores for the Australian and Mexican respondents. The levels of the nominal variable (in this case, Australian and Mexican) are represented on the horizontal x -axis, and the dependent variable values are shown on the vertical y -axis. For each group, a point is placed along the y -axis that represents the mean for the groups, and a bar is drawn to visually represent the mean value. Bar graphs are often used when the values on the x -axis are nominal categories (e.g., Australian and Mexican samples, or high-power condition and low-power condition). In contrast, line graphs are used when the values on the x -axis are numeric (e.g., number of hours that teenagers work; [Chapter 7](#)). In this case, a line is drawn connecting data points to represent the relationship between the variables.

Figure 12.5 Honest and exaggerated graphs of mean life satisfaction scores comparing two nationalities



In [Figure 12.5](#) we have also illustrated a common trick that is used to mislead readers. The trick is to exaggerate the distance between points on the y -axis to make the results appear more dramatic than they really are. Suppose, for example, that a politician wanted to mislead people into thinking that Australians are very dissatisfied with their lives compared with Mexicans (perhaps to argue in favour of some existing Mexican social program). They could use the same data in the left panel of [Figure 12.5](#) to create the graph on the right. At first glance, it appears that Australians are dramatically dissatisfied with life. But look more closely: The scale on the y -axis has been adjusted to exaggerate the difference between groups. On the left, we see the full range of the scale, from 1 to 10. On the right, the y -axis only runs from 7 to 9, making this same difference

appear much bigger. It is always wise to look carefully at the numbers on the axes for graphs. Similarly, to be an ethical researcher, you should display your results so that the total range of the scale appears on the axes of your graphs, or make it very explicit when you only present part of the scale.



LO4 Describing Effect-Size between Two Groups

It is important to describe relationships among variables in terms of size, amount, or strength. Effect-size is a general term for these indicators and is a vital descriptive tool to help us interpret how large our effects are. Effect-size can be measured in many different ways, depending on the study design. When comparing two groups (e.g., two conditions in an experiment, two nationalities) on their responses to a continuous variable, one appropriate effect-size is Cohen's *d* (Cohen, 1988). Cohen's *d* is the difference in means between two groups, standardized by expressing it in units of standard deviation (i.e., how many standard deviations large is the difference in scores between these two group?). In a true experiment, free of threats to internal validity, the Cohen's *d* value describes the magnitude of the effect of the independent variable on the dependent variable. In a study comparing naturally occurring groups, the Cohen's *d* value describes the magnitude of the effect of group membership on a continuous variable. Page 236

Because Cohen's *d* expresses effect-size in units of standard deviation, a *d* of 1.0 means that the means are 1 standard deviation apart. Accordingly, a *d* of .2 indicates that the means are separated by .2 standard deviations (or 1/5th [20 percent] of a standard deviation). The smallest possible value for Cohen's *d* is 0,

indicating no effect, but there is no maximum value. Adding confidence intervals around effect-sizes offers even more valuable information about the range of effect-sizes likely to be true (Cumming, 2012, 2014; Kline, 2013; [Chapter 7](#)).



TRY IT OUT!

To calculate Cohen's d , we simply find the difference between the two means in question, and divide this difference by the pooled standard deviation. In this way, we express the size of the difference between two means, in units of standard deviation. The pooled standard deviation is basically an average of the standard deviations for the two groups. Here is the formula for Cohen's d (we will assume equal sample sizes for both groups, for these examples):

$$d = \frac{M_1 - M_2}{SD_{pooled}}$$

And here is the formula for the pooled standard deviation:

$$SD_{pooled} = \sqrt{\frac{SD_1^2 + SD_2^2}{2}}$$

Note that for the pooled standard deviation, if standard deviations are equal for both groups, then the pooled standard deviation is equal to that same number (e.g., if both groups have a standard deviation of 2, then the pooled standard deviation is also 2). Try plugging some numbers into the formula for the pooled standard deviation to demonstrate to yourself that this is true. Then try calculating the Cohen's d effect-size for the difference between the two groups below. You can find the answer at the bottom of the page.

GROUP 1: $M = 3.30$, $SD = 1.40$

GROUP 2: $M = 4.00$, $SD = 1.40$

► Answer

Let's return to World Values Survey data comparing Australian and Mexican data for life satisfaction. The mean life satisfaction among Australian respondents was 7.20 and the standard deviation was 2.05. The mean life satisfaction among

Mexican respondents was 8.51 and the standard deviation was 1.93. Using these values and the formula in [Appendix B](#), and assuming equal sizes for the two groups for now, we can calculate that Cohen's $d = .66$. Thus, life satisfaction among Mexican respondents was .66 standard deviations higher than life satisfaction among Australians (or two-thirds of a standard deviation).

Table 12.2



TEST YOURSELF! Calculate the Cohen's d for the following sets of data, assuming equal sizes for both groups (answers at bottom of the page)

Data	Mean Difference	Pooled SD	Cohen's d
$M = 23.40, SD = 12.3$			► Answer
$M = 27.20, SD = 12.3$			
$M = 2.40, SD = 0.2$			► Answer
$M = 2.52, SD = 0.2$			
$M = 4.60, SD = 2.1$			► Answer
$M = 4.72, SD = 2.1$			
$M = 3.65, SD = 0.8$			► Answer
$M = 4.55, SD = 1.2$			



LO5 Describing Relationships among Continuous Variables: Correlating Two Variables

Different analyses are needed when you do not have distinct groups you wish to compare, but instead have a range of scores to investigate in terms of their relationship with other scores. In correlational designs, each participant is measured on at least two variables, each producing a range of numbers ([Chapter 4](#)). The appropriate analysis for this form of data involves a [correlation coefficient](#), a statistic describing whether, how, and how much two variables relate to one other. Is the relationship between these variables relatively weak or strong, positive or negative, or is there no relationship at all? Note that although a correlational analysis answers these types of questions, it cannot tell us whether there is any causal relationship between the variables ([Chapter 4](#)).

There are many different types of correlation coefficients. They are all calculated somewhat differently and their use depends on the measurement scale of the two variables being analyzed (e.g., ordinal versus interval). Because it is the most common, we will focus on the *Pearson correlation coefficient*, which is usually just called the Pearson correlation, or Pearson r . This is the correlation we use when both variables have interval or ratio-scale properties (i.e., are continuous variables).

Interpreting the Pearson r Correlation Coefficient

Recall from [Chapter 5](#) that the value of a Pearson r can range from 0.00 to ± 1.00 , with this value telling us about both the strength and the direction of the relationship. A correlation of 0.00 indicates that there is no relationship between the variables. The nearer a correlation is to 1.00 (positive or negative), the stronger the relationship. In fact, a +1.00 or -1.00 correlation is sometimes called a *perfect relationship*, because changes in the two variables follow one another in a perfect fashion. The sign of the Pearson r tells us about the direction of the relationship, whether it is a positive or a negative relationship. A correlation coefficient of $-.74$ indicates a negative relationship that is stronger than the positive relationship indicated by a coefficient of $.21$.

Because the value of the Pearson r ranges from 0.00 to ± 1.00 , it is often misinterpreted as a percentage or even a probability. Try not to make this mistake! There is no straightforward interpretation of the values for the Pearson correlation (r). An $r = .50$ does *not* mean that two variables are 50 percent overlapping, or that the two variables are related 50 percent of the time, or that there is a 50 percent chance of one variable predicting the other. We will cover how to properly interpret values of r below.[Page 238](#)

To calculate a correlation coefficient, we need to obtain pairs of observations. In other words, each individual in our study will provide two scores, one for each variable. [Table 12.3](#) shows the data for ten fictitious World Values Survey respondents for the measures of life satisfaction and subjective feelings of physical health. Respondents rated their physical health on a scale that ranged from 1 (*poor*) to 4 (*very good*) and their life satisfaction on a scale that ranged from 1 (*completely dissatisfied*) to 10 (*completely satisfied*). With these data, we can see whether the two variables are related by calculating a correlation coefficient. Are increases in one variable (e.g., physical health) accompanied by increases in the other (e.g., life satisfaction)? To find out, we will calculate a Pearson correlation coefficient (r). See [Appendix B](#) for the specific calculations. For now, we'll consider how to interpret and graph these values.

Table 12.3 Pairs of scores for subjective physical health and life satisfaction

Participant Identification Number	Subjective Physical Health	Life Satisfaction
--------------------------------------	-------------------------------	----------------------

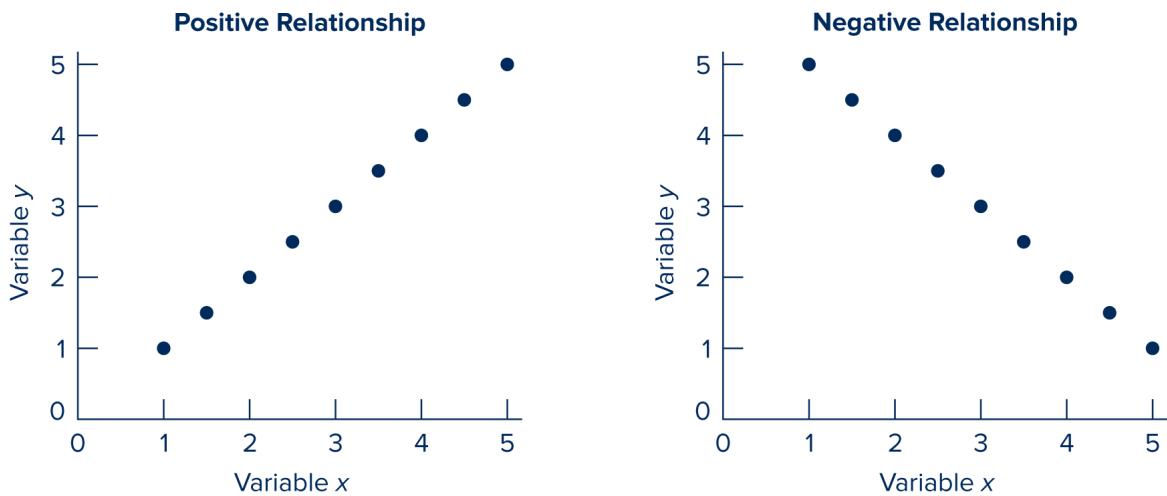
Participant Identification Number	Subjective Physical Health	Life Satisfaction
01	4	5
02	3	9
03	2	7
04	1	7
05	2	7
06	4	8
07	2	6
08	3	8
09	1	3
10	4	9

The data in [Table 12.3](#) correspond to a Pearson r value of .49, indicating a positive relationship between subjective physical health and life satisfaction. In other words, as physical health scores increase across the group, so do scores for life satisfaction. (Note that actual data from all World Values Survey respondents produced an r of .30, a smaller positive relationship.)

Scatterplots

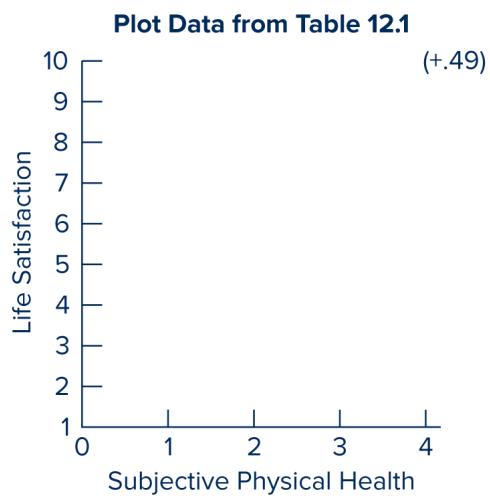
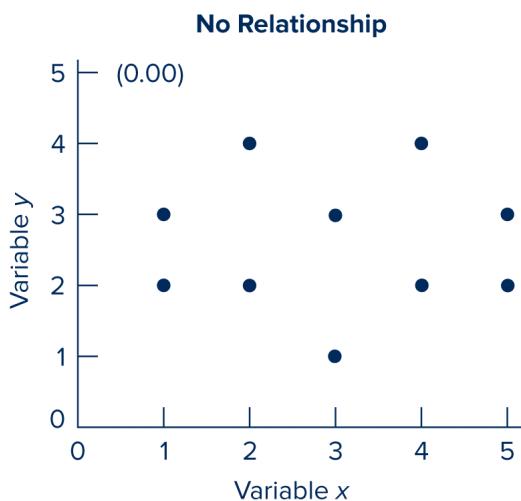
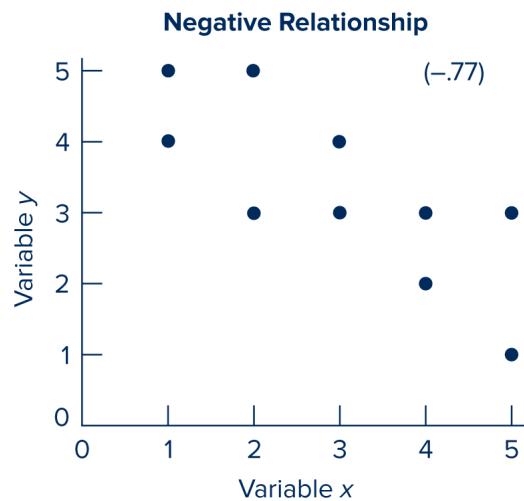
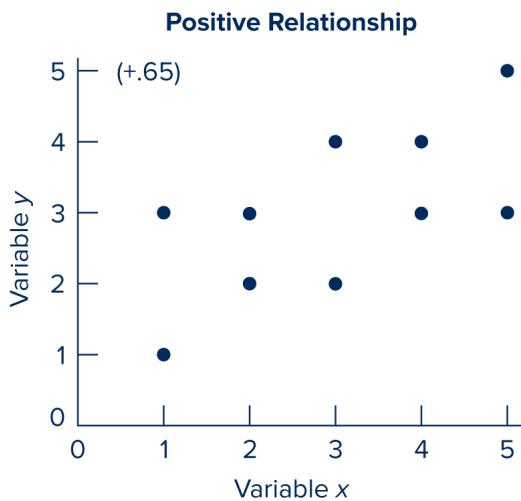
Data like those in [Table 12.3](#) can be visualized in a *scatterplot*, in which each pair of scores is plotted as a single point in a graph. [Figure 12.6](#) shows two sample scatterplots. Values of the first variable are noted on the x -axis, with values for the second variable on the y -axis. These two scatterplots show a perfect positive relationship (+1.00) and a perfect negative relationship (-1.00). You can see that for these perfect relationships, the scores fall on a straight diagonal line. Increases in one score are accompanied by perfectly predictable increases in the other score. So if we know a person's score on one variable, we can predict exactly what their score will be on the other variable. Such "perfect" relationships are rarely if ever observed in actuality. This is because all measures contain measurement error, which results in variability or differences in scores not rooted in the construct of interest. Moreover, human phenomena are almost always caused by multiple factors, so we tend not to observe just one factor predicting all of the variability in another construct.

Figure 12.6 Scatterplots of perfect (± 1.00) relationships



The scatterplots in [Figure 12.7](#) show patterns for correlations that are more realistic. You can see that the relationships depicted are not perfect relationships. The first scatterplot shows pairs of scores that are positively correlated at $.65$. From this, you can make a prediction that the higher the score on one variable, in all likelihood the score for the second variable will also be higher, but you cannot make an exact prediction. The second scatterplot shows a negative relationship, $r = -.77$. As scores on variable y increase, scores on variable x tend to decrease, but this is just a general pattern and not a perfect relationship once again.

Figure 12.7 Scatterplots depicting patterns of correlation: *Test yourself!*



Whenever relationships are not perfect, if you know a person's score on the first variable, you cannot perfectly predict what that person's score will be on the second variable. To confirm this, take a look at value 1 on variable x (the horizontal axis) in the scatterplot for the positive relationship in [Figure 12.7](#). You will see that two people had a score of 1. One of these people had a score of 1 on variable y (the vertical axis), and the other had a score of 3 on this same variable. Because the data points do not all fall perfectly on the diagonal, and there is a variation (i.e., scatter) around this diagonal line, perfect predictions are not possible. Scatterplots provide a great way of seeing how two variables relate to one another, simultaneously visualizing all of the data (i.e., not relying on statistics that summarize or average across data). These help us get a good feel for how our data actually look.

Scatterplots also allow a researcher to detect *outliers*, which are scores that are extremely distant from the rest of the data. Particularly when samples are small, outliers can skew correlation coefficients. As noted [earlier](#), sometimes outliers indicate data entry errors, and other times, these outliers might reflect real (but simply extreme) responses. In the latter case, a researcher may choose to analyze the data and report results both with and without the outlier.

The third graph in [Figure 12.7](#) (labelled “No Relationship”) shows a scatterplot for data in which there is absolutely no correlation between the two variables ($r = 0.00$). The points fall in a completely random pattern and scores on variable x are not related to scores on variable y . This is what the scatterplot would look like for the association between life satisfaction and age, from the World Values Survey data. In these data, the two variables are almost completely uncorrelated: $r = -.03$. Based on these data, age has almost no association with life satisfaction. This also means that we cannot predict how satisfied with life somebody is based on their age.



TRY IT OUT!

Try creating a scatterplot of your own. The fourth graph in [Figure 12.7](#) has been left blank so that you can plot the data found in [Table 12.3](#). The horizontal x -axis is for subjective physical health and the vertical y -axis is for life satisfaction. To complete the scatterplot, plot the ten pairs of scores on the two variables. For each individual in the sample, find the score for subjective physical health and then go up until you reach that person’s life satisfaction score. A point placed there will describe the score on both variables. There will be exactly ten points on the finished scatterplot.

Important Considerations

Restriction of Range

It is important that the researcher sample from the full range of possible scores for both variables. If the full range is not sampled, but instead this range is restricted (i.e., narrowed), the correlation coefficient produced with these data can be misleading. Examine the positive relationship depicted in [Figure 12.7](#). Imagine the people in your sample had only scores of 1 or 2 on variable x ; use your hand to cover the right side of that graph, blocking out values higher than 2 on the x -

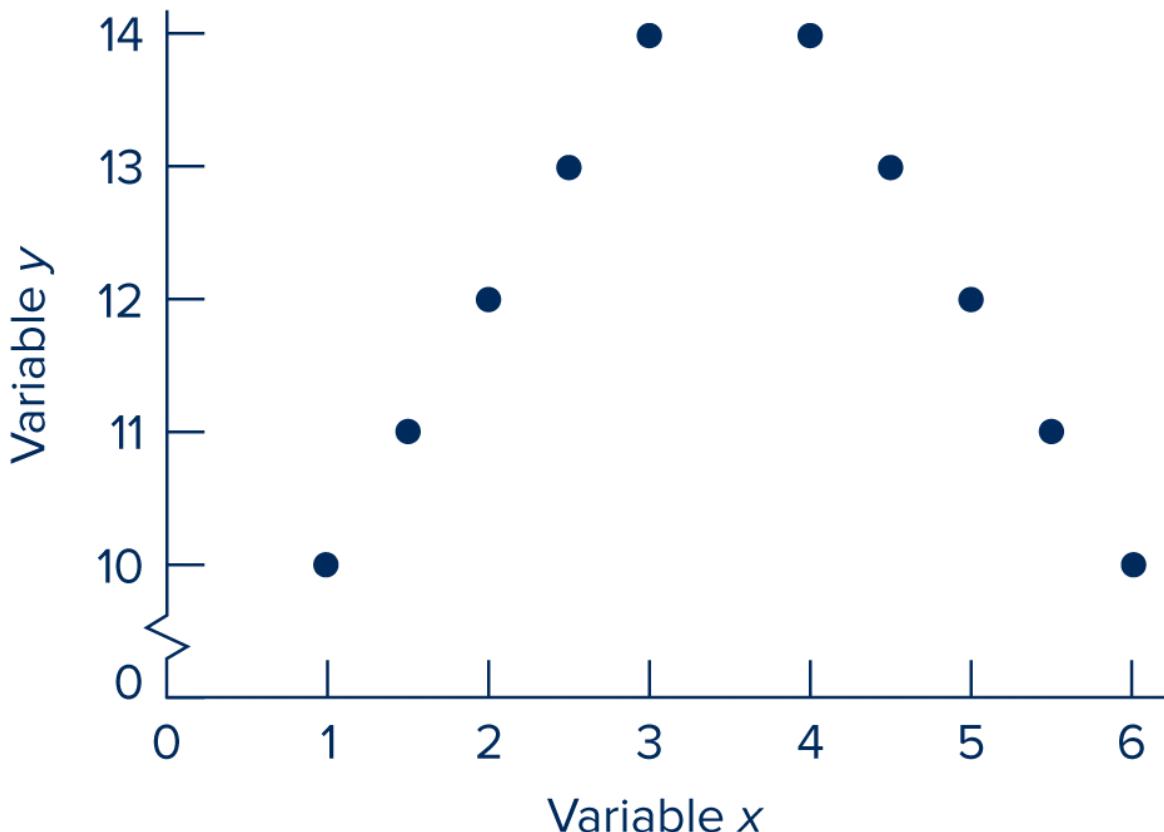
axis. Instead of showing a positive relationship, the variables now seem unrelated. With a restricted range of values, there is less variability in the scores and thus less variability that can be explained or predicted by the other variable.

The problem of *restriction of range* can occur when the people in your sample are all very similar on one or both of the variables you are studying. For example, in the World Values Survey data, age is negatively correlated with subjective feelings of health, $r = -.34$. This is as we might expect: As people get older, they tend to feel less healthy. However, if we were to only look at people younger than 30, restricting the range of age, the relationship between age and subjective feelings of health shrinks to near zero, $r = -.07$. Restriction of range can lead us to mistakenly conclude that there is no relationship between variables, because there is insufficient variability in one or more of our variables to allow us to detect how changes in one relate to changes in another.
Page 241

Curvilinear Relationship

The Pearson correlation is designed to detect only linear relationships. If the relationship is not linear but curvilinear, for example (as in the scatterplot shown in [Figure 12.8](#)), the correlation coefficient will fail to detect this relationship. In fact, the Pearson r correlation for the data in [Figure 12.8](#) is exactly 0.00, even though the two variables are clearly related. As scores on x increase so do scores on y , up until a point at which further increases in x are accompanied by decreases in y .

Figure 12.8 Scatterplot of a curvilinear relationship (Pearson correlation coefficient = 0.00)



Think about It!

Can you think of two things in real life that would have the kind of curvilinear relationship described in [Figure 12.8](#)? What is a real-world example of two variables that have a positive relationships up until a point, after which we observe a negative relationship between the two variables?

Because the variables in your data might have a curvilinear relationship, it is important to always inspect a scatterplot of the data before looking at the magnitude of the correlation. The scatterplot is valuable because it provides a visual indication of the shape of the relationship. When the relationship is curvilinear, another type of statistic must be used to determine the strength of the relationship. All forms of statistical software can produce scatterplots and can also be used to determine how well the data fit a linear or curvilinear relationship.

Correlation Coefficients as Effect-Sizes

Recall that an *effect-size* statistic is an important way to describe research results (e.g., Cohen's d). Effect-size estimates provide a scale of values that is consistent and can be compared across studies, regardless of the variables, the particular design, or the number of participants. Correlation coefficients not only allow us to examine relationships between continuous variables, they are also indicators of effect-size. But how are we to interpret correlations of different magnitudes?

What makes for a large or a small correlation? Perhaps the most intuitive way is to examine how commonly we encounter correlations of different sizes. Hemphill (2003) conducted a large survey of past research and discovered that the bottom-third of correlations observed are below .20, the middle-third ranges from .20 to .30, and the top-third is greater than .30. These empirically derived guidelines help us to interpret whether the correlation we have observed is around the range of what is typically observed, is smaller, or is larger.^{Page 242}

As an example of using correlation coefficients as effect-sizes, consider the results from studies examining eating restraint among twins. Eating restraint is the degree to which people monitor and restrict their food intake. The similarity in eating restraint scores for identical twins is equivalent to a correlation of .55, demonstrating that these pairs are highly similar in the degree to which they attempt to control their food intake. The correlation for fraternal twins, in contrast, is equivalent to an r of .31. The fact that identical twins are more similar in their eating restraint than are fraternal twins led researchers to conclude that there is a genetic component to restrained eating (Schur, Noonan, Polivy, Goldberg, & Buchwald, 2009).

As we discussed above, there is no simple interpretation for values of r . However, if we square the value of r , by multiplying it by itself, it does lend itself to a simple interpretation. This squared value of r , known as r^2 , is the proportion of variance being explained. For example, if a correlation is $r = .50$, then the $r^2 = .25$, which means that one variable explains or accounts for 25 percent of the variance in the other variable, and vice versa (remember that we do not know causal direction based on a correlation). Thus, the range of r^2 values run from 0.00 (0 percent) to 1.00 (100 percent). This r^2 value, or squared correlation coefficient, is sometimes referred to as the amount of *shared variance* between the two variables.



TRY IT OUT!

Consider the relationship between life satisfaction and subjective reports of health from the World Values Survey. The correlation coefficient is $r = .30$. Convert that value to r^2 , what value do you get? Now multiply this by 100; what does this value mean?

- Answer

Describing Relationships among Continuous Variables: Increasing Complexity

The remainder of this chapter builds on your basic understanding of correlation by discussing regression and its ability to account for additional variables. We will assume variables are continuous (i.e., measured on interval or ratio scales), although these analyses can also be adapted for use with other types of variables.



LO6 The Regression Equation

An advanced way of examining how variables relate or covary involves a statistical technique called *regression*. Like a correlation, regression analyzes relationships among variables. In fact, the calculations for correlation and regression result in the same value when there are only two variables involved. However, using a regression framework can be more powerful because it allows us to expand the analysis to include more variables. Page 243

Analyzing data using regression requires calculating a *regression equation*, which is the same as an equation for drawing a straight line. But this is not just any line: It's the line that best summarizes all of the data points. Re-examine [Figure 12.7](#). If we were to fit a regression line to the data in the top left graph, it would be angled upward, and the line would have the smallest distance possible from all

data points. In doing so, we are using a line to summarize or characterize all of the data as best as we can. If all the data points were to fall exactly on the line, the line would be a perfect summary or characterization of the data. However, this is almost never the case. Instead, we try to draw our straight line so it is as close as possible to all the data points. The regression equation that describes this line can then be used to make specific predictions. If we know the regression equation that summarizes the ability for self-reported feelings of health to predict life satisfaction, we could plug in someone's score on the former (i.e., feelings of health) and use it to predict the latter (i.e., life satisfaction). Consider a different example. In a clinical setting, if we know the regression equation summarizing the relationship between symptom severity and treatment duration, we could measure a new client's symptom severity and use it to estimate how long that client will need treatment before symptoms are reduced.

The general form of a regression equation is

$$Y = a + bX$$

where Y is the score we wish to predict (called the *criterion variable*), X is the known score (called the *predictor variable*), a is the y-intercept (a constant, where the line hits the y -axis or the value of y when $x = 0$), and b is the slope of the line (a weighting adjustment factor that is multiplied by X). The equations to find the values of a and b are beyond the scope of this book, but you can consult a statistics textbook for formulas. For now, it will be useful to just get a sense of how a regression equation can be used. Using World Values Survey data, we find that subjective feelings of health predict life satisfaction, resulting in the following regression equation:

$$Y = 4.51 + (0.81)X$$

► Answer

Thus, if we know a person's score on X (subjective feelings of health), we can insert that into the equation and predict what that person's score on Y (life satisfaction) will be.



TRY IT OUT!

Let's say someone's score for subjective feelings of health (represented by X) is 4 (out of a possible 5, indicating feeling very good about their health). Use the regression equation to estimate their life satisfaction score. Note that unless all the data points fall right on the regression line, we wouldn't expect this predicted estimate to be exactly what is truly the case (i.e., the person's actual score). However, it is our best estimate based on the prior data we have collected. For the answer, see the bottom of the page.

Multiple Correlation and Multiple Regression

Thus far we have focused on examining the relationship between only two variables at a time. Yet many different variables may be related to a given outcome. A technique called *multiple correlation* (symbolized as R , to distinguish it from Pearson r) provides the correlation between a combined set of predictor variables and a single criterion variable. Because almost any human phenomenon is likely determined by a great number of factors, accounting for many predictor variables usually permits greater accuracy of prediction than using only one predictor. In other words, the multiple correlation usually generates a stronger relation than the single correlation between any one of the predictor variables and the criterion variable. Note that the *squared multiple correlation coefficient* (R^2) is interpreted in much the same way as the squared correlation coefficient (r^2). That is, R^2 tells you the proportion of variability in the criterion variable that is accounted for by the combined set of predictor variables. Page 244

Regression is more powerful than correlation because it can be expanded to accommodate more than one predictor to predict the criterion variable. This expanded model is often called *multiple regression*. This technique allows us to examine the unique relationship between each predictor and the criterion. This is in contrast to multiple correlation, which only provides a single value for the relationship between the combined set of predictors and the criterion variable (i.e., not the relationships between each individual predictor and the criterion, as in multiple regression). (For more about the difference between multiple correlation and multiple regression, see Huberty, 2003.) For example, we may wish to use both subjective feelings of health and satisfaction with household income to predict life satisfaction. In its general form, the expanded multiple regression equation looks like this:

$$Y = a + b_1X_1 + b_2X_2 + \dots + b_nX_n$$

where Y is the criterion variable, X_1 to X_n are the predictor variables, a is the y -intercept (a constant), and b_1 to b_n are the weights that are multiplied by scores on the predictor variables. Using the World Values Survey data, the equation predicting life satisfaction with these two variables now looks like this:

$$\begin{aligned}\text{Life satisfaction} = & 2.95 + 0.54(\text{subjective feelings of health}) \\ & + 0.40(\text{satisfaction with household income})\end{aligned}$$

Just like in regression with one predictor, we could plug in someone's ratings for subjective health and income satisfaction to predict life satisfaction. You can see how both subjective health and income satisfaction are likely related to how satisfied someone is with their life, and how including both variables as predictors is going to improve our prediction of life satisfaction.

When researchers use multiple regression to study basic research topics ([Chapter 1](#)), predicting an individual's score is not a high priority. Instead, researchers are often interested in how each individual predictor relates to the outcome variable, controlling for the influence of the other predictor variables (holding those values constant). This involves a slightly different version of the multiple regression equation. Essentially, the adjusted calculations make it possible to assume that all variables are measured on the same scale. When this is done, each predictor's weight (symbolized by b) reflects the magnitude of the relationship between the criterion variable and that predictor variable, holding the other predictors constant. One way to think about this is that if we are using height and weight to predict shoe size, the values in a regression output (b) tell us how well height predicts shoe size for people of equivalent weight, and how well weight predicts shoe size for people of equivalent height.

☆ Student Spotlight: Multiple Regression ☆

Psychopathy and Machiavellianism are two highly related antisocial traits, both associated with the exploitation of others. While an undergraduate at the University of Toronto, Andrew Brankley became interested in how these two traits relate to perceptions of others, within the context of a competitive task. Andrew and Dr. Nicholas Rule examined whether trait psychopathy and Machiavellianism were associated with how threatening others were perceived, within the context of a competition task. As part of their analyses, they used a regression analysis to uncover what would be observed if additional factors were controlled, such as the emotional expression of the person being judged and the

participant's gender. You can read more about what they discovered in the journal *Personality and Individual Differences* (Brankley & Rule, 2014).

Page 245

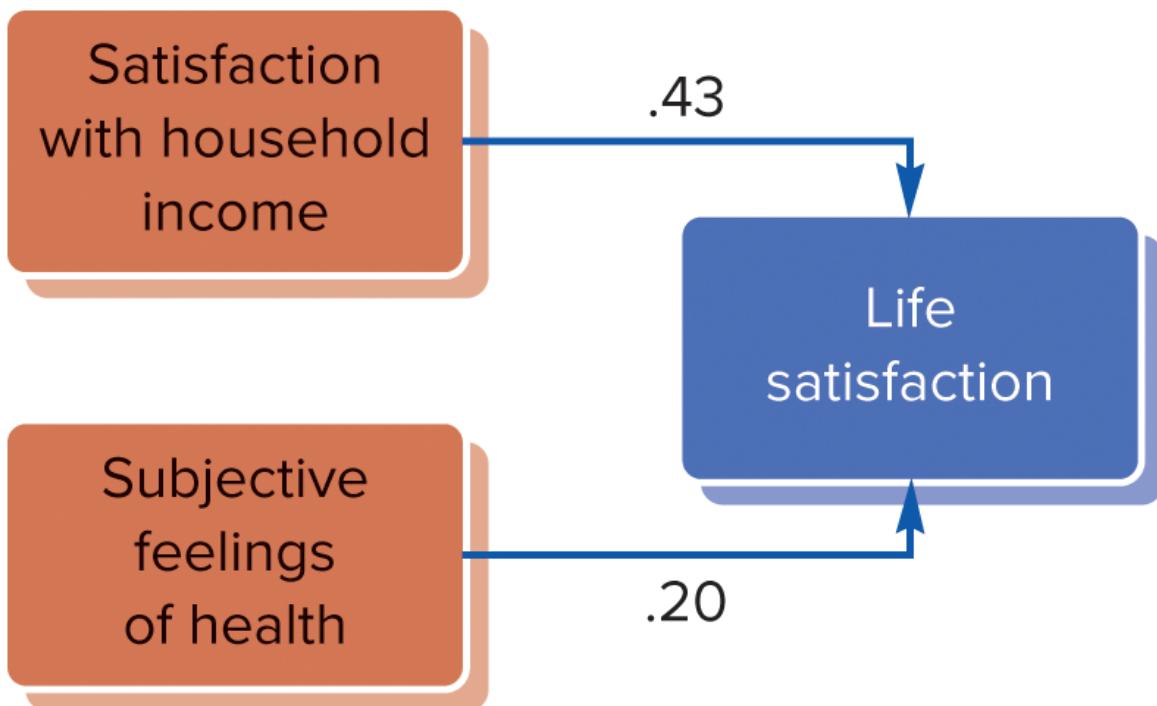
Integrating Results from Different Analyses

Let's recap what we have learned so far by combining our different analyses on the World Values Survey data. In a first step, we note that life satisfaction is correlated with each of our two intended predictors. The *correlation coefficient* between life satisfaction and subjective feelings of health is .30; the correlation between life satisfaction and satisfaction with household income is .48. Next we can calculate the *multiple correlation*. We learn that life satisfaction correlates highly with the combination of the other two variables at .52. This is equivalent to an R^2 of .27, so 27 percent of the variance in life satisfaction is being explained by income satisfaction and subjective health. Is one variable a better predictor of life satisfaction than the other? That question must be addressed using *multiple regression*. Using the two predictors of life satisfaction (i.e., subjective health and income satisfaction), the "standardized" version of the multiple regression equation is:

$$\begin{aligned}\text{Life satisfaction} = & 0.20(\text{subjective feelings of health}) \\ & + 0.43(\text{satisfaction with household income})\end{aligned}$$

Because the weight (i.e., the b value being multiplied by the predictor score) for income satisfaction is more than twice that for subjective health, we learn that life satisfaction has more to do with being satisfied with household income than with feeling healthy, but both are contributors to life satisfaction. Notice that each analysis (correlation, multiple correlation, multiple regression) provides different information about how our variables relate to one another.

You might encounter a visual depiction of regression equations in journal articles. The multiple regression model for life satisfaction that we just described could be diagrammed as follows:



LO7 Partial Correlation and the Third-Variable Problem

As we have explored, the basic correlation coefficient can be adapted in many ways to address various research questions. Another technique—called partial correlation—helps to address a specific issue: the third-variable problem. The third-variable problem occurs when two variables are correlated, but we don't know if some third variable might be the reason they are related ([Chapter 4](#)).

Partial correlation provides a way of statistically controlling for possible third variables in correlational analyses. The result of a partial correlation is a correlation between the two variables of interest, with the influence of a third variable controlled, or parcelled out of, the original correlation. It estimates what the correlation between the two primary variables would be if the third variable were held constant—in other words, if everyone responded in the same way to the third variable. This is not the same as actually keeping the variable constant, but it is a useful approximation. To calculate a partial correlation, you need to have scores on the two primary variables of interest as well as the third variable that you want to control for.^{Page 246}

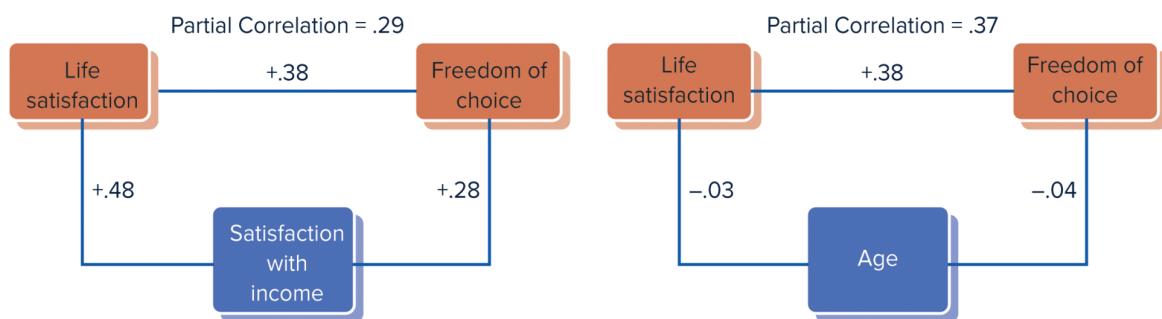
Suppose a researcher is interested in whether life satisfaction relates to perceptions of freedom of choice in their lives. After obtaining data (again from the World Values Survey), the researcher finds that the correlation coefficient is .38. However, the researcher suspects that a third variable may explain this association: satisfaction with income. Greater satisfaction with income could increase both life satisfaction and the sense that one has freedom of choice in one's life. Because satisfaction with income was also measured, it can be included in a partial correlation calculation to control for this potential influence. The resulting value will indicate the relationship between life satisfaction and perceived freedom of choice, while holding constant the effect of satisfaction with income.

When a partial correlation is calculated, you can compare the partial correlation with the original correlation to see if the third variable was influencing the original relationship. Is our original correlation of .38 substantially affected when satisfaction with income is held constant? The partial correlation controlling for income satisfaction is .29, indicating that a relationship between life satisfaction and perceived freedom remains after removing the influence of income satisfaction. Moreover, the magnitude of this association is not much different.

Figure 12.9 shows a visual depiction of two different partial correlations for the sake of comparison. The correlation coefficient between each variable is indicated by the number near the line connecting those variables. Notice that both panels show the same .38 correlation between life satisfaction and perceived freedom of choice. The first panel shows that income satisfaction correlated positively with both variables of interest. The partial correlation (removing the effect of income satisfaction) drops from .38 to .29, because income satisfaction is correlated with both of the primary variables. In the second panel, age is considered as a potential third variable. However, this partial correlation remains

almost the same at .37, because each variable is almost completely uncorrelated with age. Together, these examples show that the outcome of the partial correlation depends on the magnitude of the correlations between the third variable and both of the two primary variables.

Figure 12.9 Two partial correlations between life satisfaction and freedom of choice



Advanced Modelling Techniques

Quantitative psychologists have developed many advanced statistical techniques that are very useful, beyond the scope of introductory textbook. However, let us briefly explore one set of techniques called structural equation modelling (SEM), because it appears frequently in the research literature. SEM examines models that specify a set of relationships among many variables (Kline, 2010; Ullman, 2007). A model is an expected pattern of relationships among numerous different variables. The researcher starts by proposing a model based on a theory of how the variables might be related. Often, this model (and theory) will propose causal relations among the variables. After each variable has been measured, statistical methods are applied to examine how closely the proposed model actually “fits” the data that was obtained. Often, various models will be explored, to see which model fits the data the best. Researchers typically present path diagrams to visually represent the models being tested. Such diagrams show the causal paths among the variables. The multiple regression diagram for life satisfaction shown earlier is an example of a path diagram of a very simple model. Note, however, that although the theory and the model might propose causal relations among variables, and we can explore how well this model fits the data, SEM cannot determine whether or not a causal association exists: it is still a method based on correlation (and covariance).Page 247

There are many other applications of SEM. For example, SEM allows us to evaluate mediating variables ([Chapter 4](#)). Researchers can also use SEM to examine much more complex models that contain many more variables than with multiple regression alone (e.g., the analysis of job attitudes among women Canadian police officers by Tougas, Rinfret, Beaton, & de la Sablonnière, 2005). The major point here is that there are techniques to analyze data in complex ways, which can lead to a better understanding of the complex networks of relationship among variables.

Combining Descriptive and Inferential Statistics

The topics in this chapter have been about describing the data obtained in a study. Correlation and regression help us to describe how variables relate to one another and are part of what is known as descriptive statistics (along with measures of central tendency, measures of variability, visualizations like histograms, etc.). After describing our data, based on a sample from a population, it is common practice to try to make inferences about the population from which the sample was drawn using these sample data. These are known as inferential statistics, the topic to which we turn in the next chapter. Descriptive and inferential statistics are related, and you may find yourself returning to this chapter for some reminders about concepts that reappear (e.g., standard deviation, effect-size).

Study Terms

Test yourself! Define and generate an example of each of these key terms.

- bar graph (p. 228)
- central tendency (p. 231)
- Cohen's d (p. 236)
- correlation coefficient (p. 237)
- criterion variable (p. 243)
- descriptive statistics (p. 231)
- effect-size (p. 235)
- frequency distribution (p. 228)
- frequency polygons (p. 231)
- histogram (p. 229)
- mean (p. 229)
- median (p. 232)
- mode (p. 232)
- multiple correlation (p. 243)
- multiple regression (p. 244)
- normal distribution (p. 229)
- outliers (p. 228)

- partial correlation (p. 245)
- pie chart (p. 229)
- predictor variable (p. 243)
- regression equation (p. 242)
- restriction of range (p. 240)
- scatterplot (p. 238)
- squared correlation coefficient (p. 242)
- squared multiple correlation coefficient (p. 243)
- standard deviation (p. 230)
- variability (p. 233)
- variance (p. 233)

Review Questions

Test yourself on this chapter's learning objectives. Can you answer each of these questions?

1. What information does a frequency distribution provide?
2. Distinguish among a pie chart, bar graph, histogram, and frequency polygon. Construct one of each.
3. What information do measures of central tendency provide? Distinguish among the mean, median, and mode.
4. What information do measures of variability provide? Distinguish between the standard deviation and the range.
5. Under what circumstances would a researcher choose to compare percentages, compare means, or correlate scores?
6. What is a correlation coefficient? What do the size and sign of the correlation coefficient tell us about the relationship between variables?
7. What information does an effect-size provide? Interpret Cohen's d and correlation coefficients as effect-size indicators, and compare when each is used.
8. What is a scatterplot? What happens when a scatterplot shows a curvilinear relationship?
9. How does multiple correlation increase the accuracy of prediction over a correlation coefficient like Pearson r ?
10. What is a regression equation? How is regression similar to and different from a correlation?

11. What is the purpose of partial correlation?

Deepen Your Understanding

Develop your mastery of these concepts by considering these application questions. Compare your responses with those from other people in your study group.

1. Hill (1990) studied correlations between final exam score in an introductory sociology course and several other variables (e.g., number of absences). The following Pearson r correlations with final exam score were obtained:

Overall college GPA	.72
Number of absences	−.51
Hours spent studying on weekdays	−.11
Hours spent studying on weekends	.31

Describe each correlation and draw graphs depicting the general trend of each relationship. Why do you think grades are differently correlated with hours spent studying on weekends versus weekdays?

Page 249

2. Ask 20 students on campus to estimate how many hours per week they spend studying and how many hours per week they spend working in paid employment.
 1. Create a frequency distribution and find the mean for each of the two variables.
 2. Construct a scatterplot of these two variables. Does there appear to be a relationship between the variables? (*Note:* If there is likely to be a restriction of range problem because few students on your campus work, ask different questions, such as number of Facebook friends or hours spent watching television each week.)

3. Divide your sample into people who work in paid employment and people who do not. Calculate the mean number of hours spent studying separately for each group. Choose an appropriate graph to display your findings.
 4. What can you conclude about the relationship between these two variables? What can't you claim? Why?
3. Before the school year began, Ms. King reviewed the folders of students in her incoming Grade 4 class. For a Grade 3 reading comprehension test, which had a maximum score of 100 and a mean score of 70, she found that the standard deviation of scores was exactly 15. What information does this provide her?
4. Select the correct answer(s) to questions a, b, and c.
1. Which of the following numbers could *not* be a squared multiple correlation coefficient (R^2)?
-.99 +.71 +1.02 -.01 +.38
 2. Which one of the following correlation coefficients indicates the strongest relationship?
.23 -.89 -.10 -.91 +.77
 3. Which of the following correlation coefficients indicates the weakest negative relationship?
-.28 +.08 -.42 +.01 -.29

Inferential Statistics: Making Inferences about Populations Based on Our Samples



©DaydreamsGirl/Getty Images

Most research examines one small sample from some larger population, like asking questions of this single chick and not everyone in the brood. But do the answers we get from this one chick reflect what we would hear if we could ask every bird in the brood? Inferential statistics try to answer this difficult question.

Learning Objectives

Keep these learning objectives in mind as you read to help you identify the most critical information in this chapter.

By the end of this chapter, you should be able to:

1. LO1 Explain the purpose of using inferential statistics to evaluate sample data.
2. LO2 Distinguish between the null hypothesis and the research hypothesis.
3. LO3 Discuss how a sampling distribution is used to determine statistical significance.
4. LO4 Describe when to use the *t*-test, and explain the three basic steps to using it.
5. LO5 Describe the *F* test, including systematic variance and error variance.
6. LO6 Compare and contrast Type I and Type II errors, including elements that influence the probability of making them, and their impact on scientific progress.
7. LO7 Discuss multiple reasons statistically non-significant results may occur.
8. LO8 Discuss how effect-size and confidence intervals contribute to the validity of conclusion beyond hypothesis tests.

Page 251 What does our sample tell us about the population from which it was drawn? In the previous chapter, we examined ways of summarizing the results of a study using descriptive statistics. Researchers are also interested in inferring whether the results obtained in a particular sample represent what we would find in the entire population from which that sample was drawn. In order to do so, we use what are known as *inferential statistics*. We will also consider some of the many complex issues related to deciding whether results from a sample generalize to the population. Although we have tried to simplify this discussion whenever possible, concepts in this chapter are undoubtedly challenging. In fact, there is an active debate among researchers about many of the issues raised in this chapter. But fear not! Take your time, plan to read it more than once, complete the end-of-chapter questions, discuss the concepts with your classmates and instructor, and consider seeking out additional readings to supplement your understanding.



LO1 Inferential Statistics: Using Samples to Make Inferences about Populations

The results of any given study are typically based on a sample of participants drawn from a larger population. Researchers rarely study entire populations, because often the populations we are interested in are simply too large, making it impossible to gather data from every single member of that population. However, we want to understand and make claims about these populations of interest. Do the results from our sample appear similar to those that would be observed if we included the entire population? For example, would the difference in life satisfaction between Mexican and Australian respondents remain if we tested the whole population of each country ([Figure 12.5](#))? Inferential statistics are a way to help us infer whether a specific result observed in a sample reflects what we would observe in the population.

The most common form of inferential statistics in psychology is null-hypothesis significance testing (NHST), most readily identified as any technique that results in the calculation of a *p*-value (i.e., a probability value). However, this form of statistics has been heavily criticized,

especially in recent years (Cumming, 2012, 2014; Kline, 2013; McShane, Gal, Gelman, Robert, & Tackett, 2019). That said, learning the concepts in this chapter will help you to engage with the existing literature, which frequently relies on these types of statistics. Many researchers are calling for the field to move away from NHST and look to emphasize other methods, including effect-sizes, confidence intervals, and Bayesian statistics. Therefore, we will further explore some of these topics at the end of this chapter. However, a consensus has yet to emerge in this complex debate. As a burgeoning researcher, you should read as much about this topic as you can, and begin to form your own informed opinions about how to best make inferences from sample data. As a scientist, it is your responsibility to learn as much as you can and evaluate the available evidence to come to your own conclusion, rather than unthinkingly accept the status quo or the opinions of others.

Inferential Statistics: Ruling Out Chance

Much of our earlier discussion of experimental design centred on the importance of ensuring that groups are equivalent in every way except for the manipulation of the independent variable ([Chapter 8](#)). Groups are made equivalent by controlling all other variables and using techniques like random assignment or a within-subjects design. If the groups are equivalent except for what you have manipulated, then any differences in the dependent variable are assumed to be caused by the manipulated independent variable.[Page 252](#)

In order for this assumption to be valid, an experiment must be very well-designed and free of threats to internal validity. However, it is also true that even when an experiment has high internal validity, there will almost always be some difference between any two groups even ignoring any experimental manipulation. This happens because we are dealing with samples rather than populations. Random error will be responsible for some difference between groups, even if the independent variable had no effect on the dependent variable. Inferential statistics are a way to judge whether the difference between means reflects a real effect of the independent variable that would also be observed within the population, or simply random error. When inferential statistics conclude that the difference

reflects a real difference in the population, it is described as *statistically significant*, a term that should not be confused for what is commonly meant by the word *significant*. Statistical significance does not indicate the importance of an effect, or how meaningful it is.

Statistical Significance: An Overview

After collecting data and calculating descriptive statistics, for many researchers the next step is deciding whether those values are *statistically significant*. If we observe a difference between groups for an experimental design, we can ask, “Are these differences likely due to random error, or do they likely reflect a real effect of the experimental manipulation that would be observable in the population?” To try to answer this question, researchers can analyze their data using a statistical test. There are many different kinds of statistical tests, and which is most appropriate depends primarily on the study’s design and the type of data that results. The logic underlying the use of any statistical test rests on statistical theory, which is grounded in probability theory.

There are some general concepts that may help you understand statistical tests. (Don’t worry if you don’t understand these right away, as we will expand on them throughout this chapter.) First, the goal of any statistical test is to inform a judgment about whether an effect observed in a sample is good evidence of a real effect in the population. Second, for each test, you need to decide how willing you are to be wrong if you conclude that there is an effect in the population. This is called the *significance level* or *alpha level* (also known as the threshold for statistical significance), because it is signified by alpha (α). For NHST statistics, we calculate a probability value, or *p*-value, and then check whether this value is larger or smaller than alpha (i.e., the significance level). If it is smaller than alpha, then we describe the result as statistically significant. Third, you are most likely to obtain statistically significant results (i.e., $p < .05$) when you have a large sample size, because larger sample sizes can provide better estimates of true population values. Finally, you are more likely to obtain statistically significant results when the effect-size is large (e.g., when differences between groups are large and the variability of scores within groups is small). In the remainder of this chapter, we will expand on all of these

concepts, beginning with the null hypothesis and some basic concepts regarding probability.



Think about It!

It is very important to remember that statistical significance does not mean that something is significant in other ways. A statistically significant result is not necessarily meaningful, important, or even a large effect (Fraley & Marks, 2007). This can become quite confusing when people use the word significant to refer to statistical significance, leading readers to mistake this for other types of significance (i.e., something important or noteworthy, worthy of our attention). Try to think of ways to remind yourself, and others, that statistical significance has absolutely no relation to meaningful significance, or the word significant as we typically understand it.



LO2 Null and Research Hypotheses

Statistical inference is based around comparing a statement of the null hypothesis with a research hypothesis (also known as the alternative hypothesis). In [Chapter 2](#) we discussed the idea of a hypothesis: the researcher's tentative idea about how variables are related. In the framework of inferential statistics, this general hypothesis—that there is a relationship among variables in the population—is called the [*research hypothesis*](#) (signified by H_1). For an experiment, the research hypothesis is that the means are *not* equal in the population: They are different in some way. For a correlational study, the research hypothesis is that there is some association between two variables in the population. Notice that the hypothesis is framed in terms of the population because we are trying to infer whether our results would also be observed if we were able to gather data from everyone in the population of interest.

We also need to consider the possibility that any effect observed in our sample is due to error. This possibility is the focus of the [*null hypothesis*](#) (signified by H_0), which hypothesizes that there is no effect in the

population. In experiments, the null hypothesis states that the independent variable has no effect on the dependent variable in the population. For correlational studies, the null hypothesis states that the two variables are completely unrelated in the population ($r = 0.00$). The null hypothesis and the research hypothesis capture the only two possibilities: There either *is* or *is not* an effect in the population. Inferential statistics are used to help decide, based on our sample data, which possibility is more likely to be true in the population. Let us consider an example, using an experiment we discussed in [Chapter 4](#) that looked at the comprehension of a lecture in the context of laptop use, conducted at York University (Sana et al., 2013). This study compared people's comprehension of a lecture after sitting either (1) behind people who were multitasking on laptops or (2) behind those who were not using laptops. The null and research hypotheses for this study are as follows:

- H_0 (null hypothesis): The mean comprehension score for the Laptop group is *equal* to the mean score for the No Laptop group in the population.
- H_1 (research hypothesis): The mean comprehension score for the Laptop group is *not equal* to the mean score for the No Laptop group in the population.

The goal of null-hypothesis significance testing (NHST) is to reject the null hypothesis as unlikely, claiming that there is a very low probability that the obtained results could be due to random error. This is what is meant by statistical significance: A statistically significant result is one that has a low probability of occurring if there is no effect in the population. It is unlikely that the difference between the sample means (or the non-zero correlation) was due to random error.

For those who employ NHST, the logic is this: If we can determine, based on our data, that the null hypothesis is very unlikely to be true in the population, we reject it and conclude that our results are consistent with the research hypothesis. The null hypothesis is used because it is a very precise statement: It claims that the population means for two groups are exactly equal, or that the correlation in the population is exactly zero. This precision allows us to know the probability of the study's outcome

occurring if the null hypothesis is true in the population. Such precision isn't possible with the research hypothesis, so we infer that the research hypothesis is likely true in the population only by rejecting the null hypothesis as unlikely. Thus, statistical significance is a matter of probability.

Probability and Sampling Distributions

Probability is the likelihood or chance that some event or outcome occurs. People use probabilities all of the time in everyday life. For example, when a weather forecaster says that there is a 10 percent chance of rain today, this means that the probability of rain is quite low. If you say that you are likely to get an A+ in this course, you mean that this outcome has a high probability of occurring. Hopefully, your probability statement is based on specific information, such as your test grades for this course.

Probability in statistical inference with NHST is used in much the same way. We want to specify the probability that an event will occur if there is no difference in the population. In this case, the event we are interested in is observing a difference in group means in our sample data from an experiment. The question is: What is the probability of obtaining this result, if only random error is operating? If this probability is very low, those who use NHST reject the possibility that only random error is responsible for the difference in means, and characterize the results as statistically significant.

Probability: The Case of Mind Reading

The use of probability in statistical inference might be understood more intuitively if used in an example. Suppose that a friend claims to be able to read minds (i.e., have extrasensory perception [ESP]). You decide to test your friend by asking him to tell you which side of a coin lands face up (heads or tails) when you toss it hidden from his view. In your single case experiment, you have ten coin flip trials. Your task is to figure out whether your friend's answers reflect random error (i.e., guessing) or something more than random error. The null hypothesis in your study is that only random error is operating. The research hypothesis is that the number of correct answers is due to something more than random or chance guessing. (Note that rejecting the null hypothesis could mean that your friend has mind-reading ability, but there could also be many other

possible explanations, such as the coin was weighted toward one side, you somehow cued your friend to the correct answers, and so on.)

You can determine the number of correct answers to expect if the null hypothesis is true. Just by guessing, one out of two answers should be correct (50 percent). Therefore, on ten trials, five correct answers are expected under the null hypothesis. If, in the actual experiment, more than five correct answers are obtained, would you conclude that the obtained data reflect random error or something more than random guessing?

Suppose your friend gets six correct. You would probably conclude that only guessing is involved because you would recognize that there is a high probability of six correct answers, even though only five correct answers are expected under the null hypothesis. You expect that exactly five answers in ten trials would be correct in the long run, if you conducted this experiment with this friend over and over and over again. However, small deviations away from the expected five are highly likely in a sample of only ten trials.

Suppose, though, that your friend gets nine correct. You might conclude that the results indicate more than random error in this one sample of ten observations. You might judge intuitively that an outcome of 90 percent correct, when only 50 percent is expected, is very unlikely. At this point, you may decide to reject the null hypothesis and state that the result is statistically significant: It is unlikely to occur if the null hypothesis is correct.



LO3 Sampling Distributions

You may have been able to judge intuitively that obtaining nine correct on the ten trials is very unlikely. Fortunately, we don't have to rely on intuition to determine the probabilities of different outcomes. [Table 13.1](#) shows the probability of actually obtaining each of the possible outcomes in the mind-reading experiment with ten trials and a null hypothesis of 50 percent correct. An outcome of five correct answers has the highest probability of occurrence (0.24, or a 24 percent chance). Also, consistent with intuition, an outcome of six correct is almost as probable as getting five correct, but an outcome of nine correct is far less likely (0.0098, or less than a 1 percent chance).

Page 255

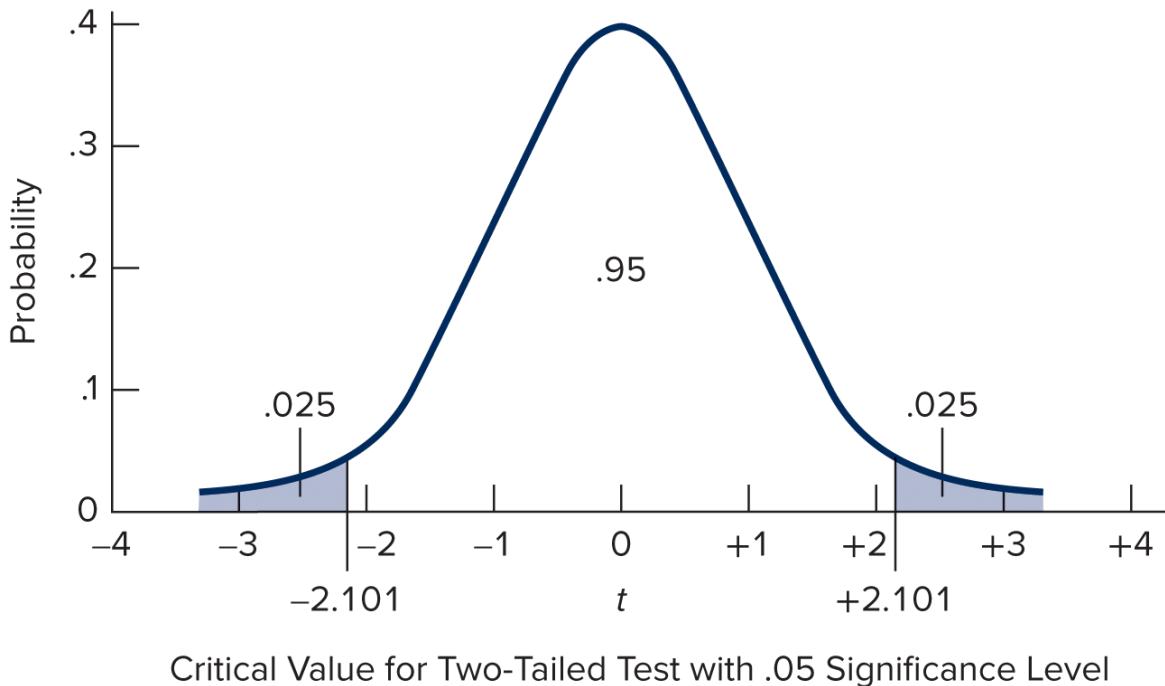
Table 13.1 Exact probability of each outcome of the mind-reading experiment

Number of Correct Answers	Probability (Based on a 50% Chance of Guessing Correctly)
10	.0010
9	.0098
8	.0439
7	.1172
6	.2051
5	.2461
4	.2051
3	.1172
2	.0439
1	.0098
0	.0010

The probabilities shown in [Table 13.1](#) were derived from a probability distribution: the outcome of a mathematical function that provides the probabilities of different possible outcomes. This particular probability distribution is called the binomial distribution, because each trial deals with two possible outcomes (heads or tails). All statistical significance decisions are based on probability distributions such as this one. Such distributions are called null hypothesis sampling distributions, or simply sampling distributions (see [Figure 13.1](#) for another example, displayed graphically). The *sampling distribution* is based on the assumption that the null hypothesis is true. In the mind-reading example, the null hypothesis is that the person is only guessing and should

therefore get 50 percent correct. This sampling distribution assumes that if the null hypothesis is true, and you were to conduct the study with the same number of observations (i.e., ten) over and over again an infinite number of times, the most frequent finding would be 50 percent correct. However, because of the random error possible in each sample, other outcomes are also bound to happen. Outcomes that are close to the expected null hypothesis value of 50 percent are very likely. Intuitively, outcomes further from the expected result (e.g., a result of nine correct in your mind-reading experiment) are less, if the null hypothesis is correct. It is important to keep in mind that for NHST (i.e., any statistic that involves calculating a p -value), the null hypothesis is always assumed to be correct.

Figure 13.1 Null hypothesis sampling distribution of t values with 18 degrees of freedom



When the obtained results are highly unlikely in the null hypothesis sampling distribution, researchers who use NHST decide to reject the null hypothesis. That is, they conclude that they have not sampled from the sampling distribution specified by the null hypothesis. Instead, they decide that the data are from a different sampling distribution, and that they can argue that the research hypothesis captures the actual sampling distribution in the population. In the case

of the mind-reading example, they would argue that they have found evidence that your friend can read minds.

Here are some basic features of a null hypothesis sampling distribution:

- Assumes the null hypothesis is actually true in the population.
- Is a frequency distribution of all possible results that would occur if a study were repeated an infinite number of times, using the same sample size, drawn from the same population (assuming the null hypothesis is true), using the same study design, and calculating the same statistic (e.g., mean difference) each time.
- Is a distribution, which means that it represents a set of possible outcomes. Even if the null hypothesis is true, the results from the various samples will vary because of random error. It is more likely that samples will vary only a little from “no effect” relative to varying a lot, but sometimes random error will cause a big deviation from “no effect.”
- “No effect” (e.g., a mean difference of zero) is the most frequently occurring result because this sampling distribution assumes that the null hypothesis is true.
- If a researcher observes a result that is far enough away from “no effect,” a researcher may choose to claim that the null hypothesis sampling distribution is unlikely to represent the truth in the population. Instead, there might be a different distribution that captures the truth in the population, and the research hypothesis represents the researcher’s best guess of what that truth is.

All statistical tests in the null hypothesis significance testing (NHST) framework rely on sampling distributions to determine the probability that the results are consistent with the null hypothesis. When the obtained data are very unlikely according to the expectations of the null hypothesis sampling distribution (usually defined as a p -value of less than .05, with alpha set at .05), the researcher using NHST typically decides to reject the null hypothesis and argues that the research hypothesis is true in the population.

Sample Size

This example of a mind-reading experiment can help to illustrate the impact of sample size—the total number of observations—on statistical significance. Suppose you had tested your friend on 100 trials instead of 10 and had observed 60 correct answers. Just as you had expected 5 correct answers based on 10 trials, you would now expect 50 of 100 answers to be correct. However, 60 out of 100 has a much lower likelihood of occurrence than 6 out of 10. This is because the more observations you have, the more likely you are to obtain an accurate estimate of true population values. As the size of your sample increases, you are more confident that your outcome actually represents the true population.

How “Unlikely” Is Enough? Choosing a STATISTICAL Significance Level (Alpha)

How unlikely does a result have to be before a researcher using NHST decides to reject the null hypothesis? A decision rule is typically adopted to decide this, prior to collecting the data. The probability required in order to declaring that a result is statistically significant is called the significance level, or *alpha level*, and it is commonly set at .05. The outcome of a statistical test is considered statistically significant when there is less than a .05 probability of obtaining the results (i.e., if the *p*-value is less than .05), *assuming that the null hypothesis is actually true* (Finch, Cumming, & Thomason, 2001). In these cases, a researcher using NHST will conclude that it is very unlikely that random error is responsible for the obtained results and choose to reject the null hypothesis.²⁵⁷

Sometimes a researcher will choose an even lower alpha value, such as .01. This means that the null hypothesis will only be rejected if there is less than 1 percent chance of obtaining results like those from the sample, or more extreme results, if the null hypothesis is actually true. The decision as to which probability to choose for an alpha threshold has to do with the consequences of being wrong, when choosing to reject the null hypothesis. We will consider this issue, known as Type I errors, *later* in the chapter. It is important to note here that there is nothing magical about a .05 or a .01 alpha level. In the null hypothesis significance testing (NHST) framework, a null hypothesis rejected at .01 does not mean that a result is “more statistically significant” than one rejected at .05. Many people have made this error of interpretation, among others, so it is important to avoid it (Kline, 2013). Remember that in NHST, the significance test is pass/fail only. To discuss the size of an effect, how big or small it is, we need to calculate and interpret an effect-size ([Chapter 12](#)).

Example Statistical Tests

After specifying the null hypothesis and the research hypothesis, as well as the alpha level, it is time to conduct a statistical test. Statistical tests use probability to help us decide whether to reject the null hypothesis. Different tests are used for different research designs. We will focus on the *t*-test and the *F* test for between-subjects experimental designs. Each test can be adapted for use with a within-subjects design ([Chapter 8](#)). The *t*-test is most commonly used to examine whether the difference in mean scores between two groups is statistically significant. In the experiment on the influence of peers multitasking on laptops for comprehension, a *t*-test was an appropriate option because the researchers wanted to know whether the mean comprehension score differed between the two groups (Sana et al., 2013). The *F* test is a more general statistical test that can be used to examine differences among three or more groups, useful for factorial designs ([Chapter 11](#)). We will also revisit the Pearson *r* correlation coefficient, which is a descriptive statistic that can be evaluated for statistical significance should we choose.



LO4 The *t*-Test: Comparing Two Means

The *t-test* is a statistic that allows you to determine whether the mean difference observed between two groups likely came from the null hypothesis sampling distribution, for that sample size. The sampling distribution of all possible values of the *t* statistic, for a sample size of 20 (10 participants per group), is shown in [Figure 13.1](#). (Note that there is a different sampling distribution of *t* for each sample size.) In this figure, the *x*-axis contains all possible outcomes of *t* that we could find in our sample, if we compare the means of two groups and the null hypothesis is true in the population. Regardless of sample size, the sampling distribution of *t* has a mean of 0 and a standard deviation of 1. Note that the sampling distribution of *t* assumes that there is no difference in the population means. Thus, the expected value of *t* under the null hypothesis is zero. (Notice in [Figure 13.1](#) that zero has the highest probability of occurring.) The larger the value of *t*, the farther it is from zero and the more unlikely it is according to the null hypothesis sampling distribution. In this way, larger *t* values are less likely under this sampling distribution, and so it could be argued that observing large *t* values makes it unlikely that these values originate from the null hypothesis sampling distribution. Conceptually, the *t* value is a ratio of two aspects of the data: (1) the difference between the group means and (2) the variability within groups. The ratio may be described as follows:

$$t = \frac{\text{group difference}}{\text{within-group variability}}$$

The numerator, or top half of this equation, is group difference, which refers to the difference between your obtained means. If the null hypothesis is true in the population, you expect this difference to be zero. Mathematically, the value of *t* increases as the difference between your observed sample means increases.

The denominator of the *t* formula, or bottom half of the above equation, is essentially an indicator of the amount of error in your dependent variable. Within-group variability captures the amount of variability of scores around each mean. Recall from [Chapter 12](#) that the standard deviation (*s*) and variance (*s*²) are indicators of how much scores deviate from the group mean. The smaller the standard deviation in each group, the less error is

influencing scores. Within the t statistic formula, smaller standard deviations mean a smaller denominator, which makes the total t value larger. Larger t values are associated with smaller probabilities (i.e., smaller p -values), making it more likely that your p -value will be smaller than alpha, allowing you to reject the null hypothesis.

To use this sampling distribution to evaluate our data, we need to follow three steps. First, calculate a value of t from the obtained data. Second, identify the critical value of t that reflects our chosen alpha level, based on the null hypothesis sampling distribution for our sample size. Third, compare the obtained t to the critical value of t . If the obtained t exceeds the critical value, it has a low probability of occurring (less than .05) if the null hypothesis is true, and therefore the null hypothesis is rejected. If the obtained t is smaller than the critical value, we retain the null hypothesis. Such a t value is likely if chance is the only reason our means differ, rather than a true effect. Researchers used to manually look up critical values for t values in tables (such as [Table C.2](#) in Appendix C; for details on this procedure, see Appendices B and C), but now statistical software is used to calculate the exact probability of obtaining a particular t value, assuming the null hypothesis is true. This probability is known as a p -value, and researchers examine if this p -value is below the chosen alpha threshold (e.g., smaller than .05). If it is, the result is deemed statistically significant, and the null hypothesis is rejected.

In the results sections of journal articles, you will often see statistical tests reported in the following format: $t(36) = 4.37, p < .05$. These statements appear at the end of sentences describing the effect and provide the reader with four important pieces of information: The t test was used, with 36 degrees of freedom, resulting in an obtained t value of 4.37, which corresponds to a p -value of less than .05 (the alpha level) if the null hypothesis is true.

The p -value indicates the probability of obtaining a result at least as extreme as that observed if the null hypothesis is true. Currently, it is recommended that this value be reported precisely down to a reasonable value (e.g., $p = .026, p = .24$), resorting to the use of the less than sign ($<$) only when very small values are observed (e.g., $p < .001$) (Wilkinson and

the Task Force on Statistical Inference, 1999). Specifically, p refers to the proportion of the null hypothesis sampling distribution that is more extreme than the obtained t value. Remember that NHST statistical tests are pass/fail. We can ask only whether the p value is larger or smaller than alpha. A small p value cannot be interpreted as an indicator of the size of an effect (i.e., how large or small it is). This is also because p values are influenced not only by the size of an effect, but also the size of the sample: the larger the sample size, the smaller the p value. Page 259

The degrees of freedom are adjustments used in statistical analyses to account for the fact that we are estimating population values (e.g., group means) using data from small samples. The smaller our sample, the larger the impact degrees of freedom have on our critical value. Mathematically, degrees of freedom are the number of scores free to vary once the means are known. For example, if the mean of a group is 6.0 and there are five scores in the group, there are 4 degrees of freedom; once you have any four scores, the fifth score is known because the mean must remain 6.0. When comparing two means, the degrees of freedom are equal to $(n_1 + n_2 - 2)$, or the total number of participants in the study minus the number of groups. In the laptop experiment, the degrees of freedom were calculated as follows: $19 + 19 - 2 = 36$.



LO5 The F Test: Used When Comparing Three or More Group Means

The t -test is limited to two-group designs. For examining differences among three or more groups, researchers using an NHST framework may choose to perform an analysis of variance (ANOVA), or [*F test*](#). This is an extension of the t -test that can be used in many more research designs. When a study has only one independent variable and examines two groups, the values of F and t are virtually identical (i.e., the value of F equals t^2). However, ANOVA can be used when there are more than two levels of an independent variable or when a factorial design with two or more independent variables has been used ([Chapter 11](#)). In [Appendix B](#) you can find the calculations necessary to conduct an F test.

The F statistic is a ratio of two types of variance (hence the term, analysis of variance) that parallels the t ratio. [*Systematic variance*](#)—the numerator—is the deviation of the group means from the grand mean, which is the mean score of all participants across all conditions in the study. Systematic variance is small when the difference between group means is small, and increases as the differences in group means increases. [*Error variance*](#)—the denominator—captures how much individual scores in each group deviate from their respective group means. Because systematic variance is the variability of scores between groups, it is sometimes called *between-group variance*. Likewise, because error variance is the variability of scores within groups, it is sometimes called *within-group variance*. The larger the F ratio, the greater the group differences are relative to the amount of error, and the more likely it is that the results are statistically significant. After this has been determined, follow-up tests are required to determine which group means differ from each other (e.g., using a simple main effect analysis; [Chapter 11](#)).

Statistical Significance of a Pearson r Correlation Coefficient

The Pearson r correlation coefficient is used to describe the strength of the relationship between two variables when both variables are continuous ([Chapter 12](#)). If we wish, we can take this descriptive statistic and perform an inferential statistical test on it, using the NHST framework. The null hypothesis in this case is that the true population correlation is exactly 0.00

—the two variables are not related at all. What if you were to observe a correlation of .27? A statistical significance test will allow you to decide whether to reject the null hypothesis and conclude that the true population correlation is different from 0.00. This is done by performing a different version of the *t*-test that compares the obtained correlation coefficient with the null hypothesis correlation of 0.00. The procedures for calculating a Pearson *r* and determining whether it is statistically significant are provided in [Appendix B](#).

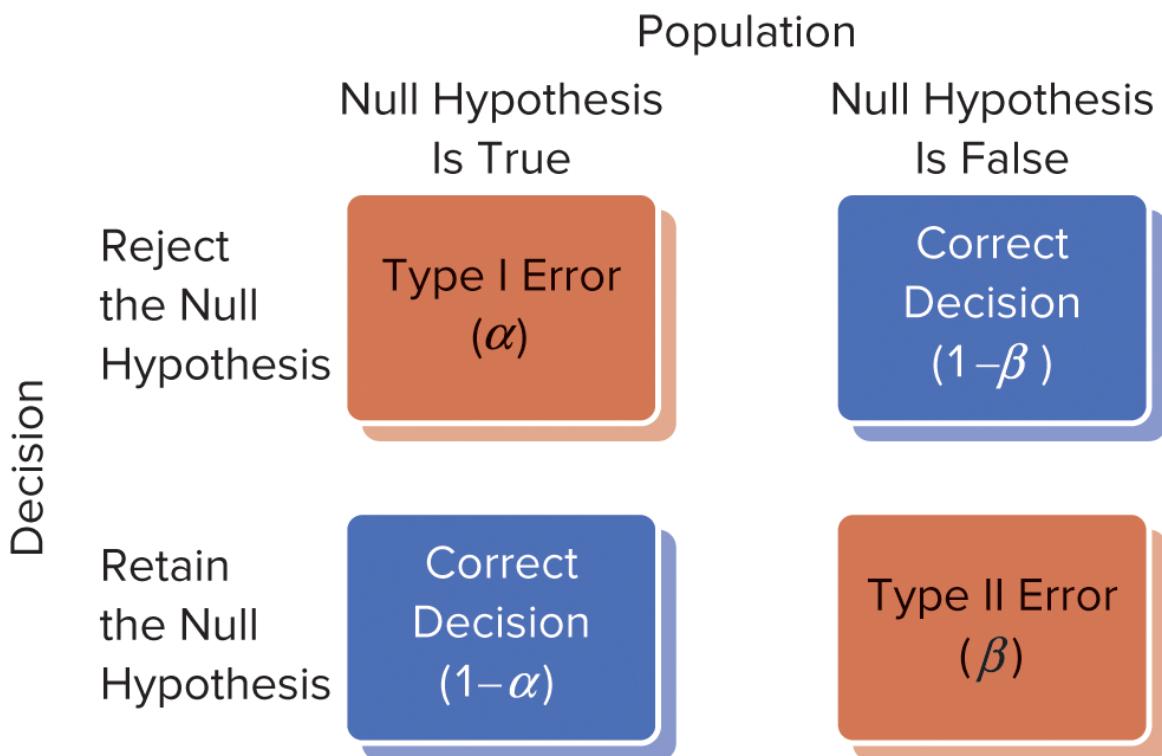


LO6 We Made a Decision about the Null Hypothesis, but We Might Be Wrong! Investigating Type I and Type II Errors

The decision to reject the null hypothesis using NHST is based on probabilities rather than certainties. We make this decision without ever knowing the truth in the population, because we are basing our judgment on just a small sample drawn from this population. Unsurprisingly, then, the decision we make might not be correct.

The ways in which we can be wrong can be described using a decision matrix ([Figure 13.2](#)). The two horizontal rows show the two possible decisions we make based on a statistical test: (1) Reject the null hypothesis or (2) retain the null hypothesis. The two vertical columns show the two possible truths about the population: (1) The null hypothesis is true or (2) the null hypothesis is false. This results in two kinds of correct decisions and two kinds of errors.

Figure 13.2 Decision matrix for Type I and Type II errors



Correct Decisions

One correct decision occurs when we reject the null hypothesis, based on our sample data, and the null hypothesis is actually false in the population. We decide that the population means are not equal or that there is some non-zero correlation between variables, and that is actually true in the population. This is the decision we hope to make when we begin our study.

The other correct decision is to retain the null hypothesis, when the null hypothesis is true in the population. In this case, the population means are in fact equal, or there is truly no relationship between the variables. There was no effect to find, and we did not accidentally find one in our sample, or accidentally conclude that an effect exists based on what we observed in our sample.

Type I Errors

A Type I error is made when we reject the null hypothesis but the null hypothesis is actually true. We decide that the population means are not equal when they

actually are equal. Type I errors occur when, simply by chance, we obtain a large value of t or F . For example, even though a t value of 4.37 is highly improbable if the population means are indeed equal, this can still happen. When we do obtain such a t value, simply by chance, we incorrectly decide that the independent variable had an effect on the dependent variable. One way to remember the Type I error is to remind yourself that this is a false positive: We falsely conclude that there is something present (i.e., a positive result), when there is nothing there in truth.^{Page 261}

The NHST framework associates the probability of making a Type I error with the choice of alpha level. When the alpha level for deciding whether to reject the null hypothesis is .05, the probability of making a Type I error (alpha) is interpreted to be .05, if the null hypothesis is in fact true. The probability of making a Type I error can be changed by either decreasing or increasing the significance level. If we use a lower alpha level of .01, for example, there is less chance of making a Type I error, assuming that the null hypothesis is true.

Type II Errors

A [Type II error](#) occurs when the null hypothesis is retained based on our sample data, but in the population the null hypothesis is actually false. In this case, the population means are not equal, or the population correlation is not zero, but the study results do not lead to a decision to reject the null hypothesis. One way to think about a Type II error is that it represents a false negative: claiming that there is nothing there (a negative result), when in truth there is something present (i.e., an effect).

Research should be designed so that the probability of a Type II error is relatively low, presuming an effect actually exists in the population. The probability of making a Type II error is called beta (β), and is related to three elements. The first is the alpha level (threshold for statistical significance). If we set a very low alpha level to decrease the chances of a Type I error, we increase the chances of a Type II error. If we make it very difficult to reject the null hypothesis, the probability of incorrectly retaining the null hypothesis increases. The second factor is sample size. True differences that exist in the population are more likely to be detected as the sample size increases. The third factor is effect-size. As the population effect-size increases, the likelihood of making a Type II error decreases. A small effect-size may go undetected, especially with a small sample. These three elements reappear in our discussion of [power](#), which is directly related to the Type II error rate.

Dr. Dan Olner (Sheffield Methods Institute) has proposed a handy way of remembering the difference between a Type I and Type II error, and which error is which. Consider the old fable of the shepherd boy who cried “Wolf!”

Originally, to amuse himself, the boy cried out “Wolf!” when there was no wolf, and all of the townspeople came running. But later, when there actually was a wolf terrorizing the sheep, the boy cried out “Wolf!” and no one came running because they did not believe that there was a wolf. This fable describes a Type I error and a Type II error, in that order. The first time the boy cried wolf, the people believed that there was a wolf present (i.e., an effect), when in truth there was none. The second time the boy cried wolf, the people believed there was no wolf, when in actuality there was a wolf (i.e., an effect).



Think about It!

This is just one way to remember what is a Type I error and what is a Type II error. Can you think of another mnemonic (i.e., a memory device) that will help you remember which type of error is which? Discuss this with your classmates and share your ideas.

The Everyday Context of Type I and Type II Errors

The decision matrix used in statistical analyses can be applied to the kinds of decisions people frequently make in everyday life. For example, consider the decision made by a juror in a criminal trial. As with inferential statistics, a decision must be made on the basis of evidence. Jurors must decide whether the defendant is innocent or guilty, but we never really know the truth.^{Page 262}

The juror’s decision matrix is illustrated in [Figure 13.3](#). To align this example to statistical decisions, assume as the null hypothesis that the defendant is innocent (adopting the dictum that a person is innocent until proven guilty). Thus, rejecting the null hypothesis means deciding to declare the defendant guilty, and retaining the null hypothesis means deciding to declare the defendant innocent. The null hypothesis may actually be true or false. There are two kinds of correct decisions and two kinds of errors, just like in statistical decisions. A Type I error is finding the defendant guilty when the person really is innocent: a wrongful conviction. A Type II error is finding the defendant innocent when the person actually is guilty: a wrongful exoneration. In North America, Type I errors by jurors generally are considered to be more serious than Type II errors. Before

finding someone guilty, the juror is asked to make sure that the person is guilty “beyond a reasonable doubt” or to consider that “it is better to have a hundred guilty persons go free (i.e., a Type II error) than to find one innocent person guilty (i.e., a Type I error).” Following this logic, to reduce Type I errors, a juror may interpret ambiguous evidence in favour of the defendant’s innocence.

Figure 13.3 Decision matrix for a juror

		True State	
		Null Is True (Innocent)	Null Is False (Guilty)
Decision	Reject Null (Find Guilty)	Type I Error	Correct Decision
	Retain Null (Find Innocent)	Correct Decision	Type II Error



The decision that a doctor makes to operate or not operate on a patient provides another illustration of how a decision matrix works ([Figure 13.4](#)). Here, the null hypothesis is that no operation is necessary. The decision is whether to reject the null hypothesis and perform the operation, or to retain the null hypothesis and not operate. In reality, the doctor is faced with two possibilities: Either the operation is unnecessary (the null hypothesis is true), or the patient will die without the operation (a dramatic case of the null hypothesis being false). Which error is more serious in this case? Doctors may believe that *not* operating on a patient who really needs the operation—making a Type II error—is more serious than making the Type I error of operating on someone who does not really need it. Following this logic, to avoid a Type II error, a doctor may choose to interpret

ambiguous symptoms as evidence that the patient does need the operation. These two examples demonstrate how Type I and Type II errors relate to one another, and how the consequences of making either kind of error affect which ones we see as more damaging, which in turn affect our decision-making. Page 263

Figure 13.4 Decision matrix for a doctor

		True State	
		Null Is True (No Operation Needed)	Null Is False (Operation Is Needed)
Decision	Reject Null (Operate on Patient)	Type I Error	Correct Decision
	Retain Null (Don't Operate)	Correct Decision	Type II Error



TRY IT OUT!

Now consider the important decision to marry someone. If the null hypothesis is that the person is “wrong” for you, and the true state is that the person is either “wrong” or “right,” you must decide whether to marry the person. Try to construct a decision matrix for this particular problem. Which error do you think is more costly: a Type I error (marrying someone who is wrong for you) or a Type II error (failing to marry someone who is right for you)?

Type I and Type II Errors in the Published Research Literature

Researchers have generally believed that the consequences of making a Type I error are more serious than those associated with a Type II error. If we conclude that there is some effect, but in truth no effect exists in the population, then a researcher might publish these results in a journal and these results might be reported in textbooks or in the media. Researchers don't want to mislead people as to what is true in the world, or risk damaging their own reputations by publishing results that are false and so cannot be replicated. The costs of arguing for something that is false, for an effect that doesn't truly exist, is high. One way to guard against the possibility of making a Type I error is to use a very low alpha level (.05 or .01). However, analyses of multiple scientific disciplines has demonstrated that setting low alpha levels alone is not enough to ensure that false "effects" are not published (Ioannidis, 2005; Simmons, Nelson, & Simonsohn, 2011).

Publication Bias Inflates Overall Type I Error Rates

One of the problems made worse by Type I errors is *publication bias*: the bias that emerges due to the fact that statistically significant results are far more likely to be published than statistically non-significant results (Kühberger, Fritz, & Scherndl, 2014). Thus, the probability of false "effects" making it into the literature is likely higher than 5 percent, which is what one would assume based on the common alpha level of .05. Publication bias has created a problem for scientists, who experience pressure to seek out statistical significance rather than truth, since statistical significance is what garners publications. Scientific productivity is often measured in terms of the number of publications, with more publications leading to other benefits such as jobs, tenure, and awards. This pressure has made commonplace some problematic research practices that increase the likelihood of Type I errors (e.g., relying on small sample sizes, selective reporting of results or analyses; see Bakker, van Kijk, & Wicherts, 2012; Simonsohn, Nelson, & Simmons, 2014). However, these Type I errors are serious and decrease the overall accuracy of our science.

Unreported Type II Errors Reduce Long-Term Accuracy

The consequences of a Type II error traditionally have not been considered as serious as Type I errors. It was thought that, at worst, a researcher gives up on

finding an effect that actually is present in the population. Recent concern over publication bias has ignited discussion of what is known as “the file drawer,” a term that describes the place where failed studies go to die. The file drawer becomes a home for a study that does not find statistically significant results, regardless of whether the researchers have accurately found no effect or have made a Type II error. These studies can easily be excluded from meta-analyses, studies that aggregate past results to describe an overall effect ([Chapter 14](#)). When failed studies are not included, meta-analyses risk overestimating the overall effect-size. It is tremendously difficult to interpret why a study failed to find a statistically significant result. One reason is that there truly is no effect to be found. But there are also many reasons why Type II errors occur (see the upcoming section on [interpreting non-significant results](#)). Many Type II errors go unknown and unreported in file drawers because of publication bias, making it difficult to build a cumulative science.

Page 264

Methodological Reform

As mentioned in [Chapter 3](#), psychologists are developing ways to improve the accuracy of our science. Efforts to reduce Type I errors have emphasized the importance of replicating findings, increasing sample sizes whenever possible, and fully disclosing measures and analyses so that others can evaluate and verify them (e.g., Eich, 2014). Websites such as the Open Science Framework (<https://osf.io/>) and AsPredicted (<https://aspredicted.org>) have been created for researchers to register research methods and data analysis plans before data collection begins, to avoid questionable research practices. In addition, sites such as [PsychDisclosure.org](#) (LeBel et al., 2013) and the Loss of Confidence Project (<https://lossofconfidence.com>) allow researchers to add information about published papers, including indicating when they no longer believe the results of a published paper should be trusted. In addition, journals have emerged that allow researchers to publish results that do not achieve statistical significance and therefore are not traditionally publishable, such as the [Journal of Articles in Support of the Null Hypothesis](#). More traditional journals have also become more open to publishing results that do not show statistical significance, especially if they provide other statistics to help establish that the data provide evidence in favour of the null hypothesis. Rather than vanishing completely, these journals allow for such studies to be available for other researchers to learn from. Importantly, these studies also become readily available for inclusion in meta-analyses. These and other reforms are striving to build a cumulative science that increases ethical practices ([Chapter 3](#)), while minimizing both Type I and Type II errors.



LO7 Interpreting Statistically Non-significant Results

Although “retaining the null hypothesis” is convenient terminology, it is important to recognize that researchers are not generally interested in retaining the null hypothesis. Research is designed to show that a relationship between variables does exist, not to demonstrate that variables are unrelated. As mentioned [above](#), statistically non-significant results are difficult to interpret. For this reason, researchers often say that they “fail to reject” the null hypothesis.

When the results of a single study are statistically non-significant, one possibility is that there is no effect in the population. Yet, many other alternative explanations are also possible. A Type II error might have occurred, in which the researcher concludes that there is no effect when there actually is an effect in the population. These Type II errors can occur because of the particular procedures used in the study. Assuming an effect really exists in the population, the results of a study may be statistically non-significant if unclear instructions are given to the participants, if the manipulation of the independent variable is too weak, or if the dependent measure is too unreliable or too insensitive ([Chapter 9](#)). In these cases, a

more carefully conducted study might be able to find the effect under investigation. Page 265

There are also statistical reasons for Type II errors. Recall that the probability of a Type II error is influenced by the alpha level, sample size, and effect-size. If the researcher chooses to use a very cautious threshold for statistical significance of $\alpha = .001$ rather than $\alpha = .05$, this would make it more difficult to reject the null hypothesis (and also more difficult to make a Type I error, if the null hypothesis is true). However, that also means that there is a greater chance of retaining an incorrect null hypothesis: A Type II error is more likely if the null hypothesis is false. In other words, a real effect is more likely to be overlooked when the alpha level is set very low.

Type II errors may also result from using a sample that is too small to detect the actual effect. Assuming a real effect exists, larger sample sizes are generally more likely to detect it. Small effects are especially difficult to detect without very large sample sizes. Generally, larger effects in the population are easier to detect than smaller ones. Consider this metaphor: Imagine you are trying to find a tiny needle in a giant haystack. Now imagine trying to find a full-sized tractor in that same haystack. Both “effects” (i.e., the needle and the tractor) are in the haystack, but it is much easier to find the much larger tractor.

Given the many causes of Type II errors, researchers should not retain a null hypothesis just because the results from one study are statistically non-significant. However, sometimes it is correct to retain the null hypothesis and conclude that two variables are *not* related. Several criteria can be used when deciding to retain the null hypothesis (Frick, 1995). Studies must be well-designed with sensitive dependent measures, and a manipulation check should indicate that the independent variable was indeed manipulated effectively. A large sample should also be used to rule out the possibility that the sample was too small to detect the effect. Further, evidence should come from multiple studies. Lastly, the NHST framework and all related statistics (i.e., any statistic that relies on calculating a p -value, including t -tests and F -tests) cannot be used to evaluate evidence in favour of the null, since these statistics assume that the null is true as a

foundational premise. Other forms of statistics, such as Bayesian statistics and different forms of equivalence testing, are required in order to evaluate evidence of a null effect. Under such circumstances, a researcher may be justified in concluding that there is in fact no relationship. For example, a Canadian graduate student, Ashley Chung-Fat-Yim, working with two undergraduates, Elena Cilento and Ewelina Piotrowska, in addition to myself (RAM), examined story engagement in bilinguals. Specifically, they wondered whether people can become just as engaged with a story when processing it in their non-native language, compared to those using their native language. Across three different studies, which presented stories in three different formats (i.e., text, podcast, and short film), they found no evidence that those using their non-native language were any less engaged in the story compared to those using their native tongue. Bayesian statistics were used to evaluate whether the data observed were more consistent with the null hypothesis (no difference in engagement between groups) or the alternative hypothesis (some difference between groups), with the data from all three studies supporting the null (Chung-Fat-Yim, Cilento, Piotrowska, & Mar, 2019).

Choosing a Sample Size: Power Analysis

We noted in [Chapter 9](#) that researchers often select a sample size based on what is typical in a particular area of research. A more appropriate approach is to select a sample size on the basis of a power analysis, before you begin any data collection. The *power* of a statistical test is the probability of correctly rejecting the null hypothesis, presuming it is actually false (Wilkinson et al., 1999). It is directly related to the likelihood that you will retain a null hypothesis that is actually false, as seen by its formula:

$$\text{Power} = 1 - p(\text{Type II error})$$

where $p(\text{Type II error})$ means “the probability of making a Type II error.” Assuming the effect really exists, a power of .80 means that you will find an effect (i.e., correctly reject the null hypothesis) 80 percent of the time, but 20 percent of the time you will miss it (i.e., make a Type II error).

As noted [earlier](#), the probability of a Type II error is related to the alpha level (threshold for statistical significance), sample size, and effect-size. A power analysis enables a researcher to choose a power probability—for example, .80—and calculate the sample size needed to detect an effect of the expected size with that probability (Cohen, 1988). [Table 13.2](#) shows the total sample size needed for an experiment with two groups and an alpha level of .05. In the table, expected effect-sizes range from $d = .10$ to $d = 1.00$, and the desired power is shown at .80 and .90. You can see from this table that smaller effect-sizes require larger samples. If a researcher is studying a relationship with an estimated effect-size of .20, a very large sample size is needed for statistical significance at the .05 level. An inappropriately low sample size in this situation is likely to produce a statistically non-significant finding, even though an effect actually exists.

Multiple reviews have found that the average effect-size observed within psychology is around $d = .40$, with a large standard deviation of around $d = .35$ (Fraley & Marks, 2007; Hemphill, 2003; Richard, Bond, Jr., & Stokes-Zoota, 2003). If you are not sure how large the effect is that you are studying, beginning with the average or something smaller is a good place to start.

Table 13.2 Total sample size needed to detect a statistically significant difference for a t -test

Population Effect-Size (d)	Power = .80	Power = .90
.10	3142	4205
.20	787	1053
.30	351	469
.40	198	265
.50	128	170
.75	58	77
1.00	34	44

Note: Effect-sizes are Cohen's d (Chapter 12). Sample sizes are based on *a priori* power analysis for a two-group between-subjects design using Cohen's d and a significance level of .05. This is the total sample size needed; for the sample size per group, simply divide by 2.

Higher desired power demands a larger sample size. This will make it more likely that you will detect any effect that is there. Researchers usually use a power of .80 or .90 when using this method to determine sample size. Several computer programs have been developed for researchers to conduct power analysis easily and determine the sample size needed for their study. See the free software program G*Power 3 (Faul, Erdfelder, Lang, & Buchner, 2007) or the "pwr" package for the statistical program R (Champely, 2018).

Analyzing Data Using Statistics Software

Although you can calculate statistics with a calculator using formulas (see [Appendix B](#)), most researchers use software to analyze their data.

Sophisticated software packages dedicated to the calculation of statistics make it easy to calculate descriptive and inferential statistics quickly and accurately. These same programs also help to visualize data by producing graphs, charts, and plots.

Some of the major statistical software programs used within the behavioural sciences are SPSS, SAS, and Matlab. In addition, the free and open source software program called R, often used in conjunction with a graphics-based interface called R Studio, is becoming increasingly popular (R Core Team, 2018). Spreadsheet programs such as Microsoft Excel can also be used for some simple analyses. The general procedures for doing analyses are similar across different statistics programs, but the specific steps will differ depending on what program you are using. In general, an analysis begins with inputting the data, running the test, and then interpreting the output.



TRY IT OUT! SOME RESOURCES FOR LEARNING R

- R Project (www.r-project.org): Your source for downloading R.
- RStudio (www.rstudio.com): A free program that allows you to use R by way of a user-friendly graphical interface.
- R for Cats (www.rforcats.net): A fun and gentle introduction to using R.

- Swirl (swirlstats.com): A way to learn R, from within R itself.
- The Personality Project's Introduction to R (personality-project.org/r/): A more detailed guide to using R.

Inputting your raw data (i.e., unprocessed data) is the first step of data analysis using any program. Suppose you want to input the data for the experiment on multitasking with laptops. You should think of your data as a matrix with horizontal rows and vertical columns. Typically, each row will represent an individual participant, and each column will represent a score on an item. If you are working in Excel, it is usually easiest to set up a separate column for each group, as shown in [Figure 13.5](#). However, other programs require different methods of data organization, with all participants from all groups in one column, and a separate column indicating which group they were in, for example.

Figure 13.5 Sample computer input and output

	A	B	C	D
1	View of Multitasking Peers Condition		No View of Multitasking Peers Condition	
2	Participant #	Proportion Correct		Participant # Proportion Correct
3	3	.4	1	.9
4	5	.65	2	.5
5	6	.575	4	.525
6	10	.625	7	.65
7	14	.325	8	.825
8	16	.45	9	.7
9	17	.625	11	.675
10	18	.625	12	.65
11	19	.45	13	.9
12	20	.55	15	.7
13	21	.725	23	.725

14 22	.6	25	.775
15 24	.675	28	.75
16 26	.8	29	.7
17 27	.55	31	.925
18 30	.55	32	.825
19 34	.425	33	.85
20 35	.475	37	.675
21 36	.5	38	.7

Excel method of data input

t Test: Two-Sample Assuming Equal Variances

	View	No View
Mean	0.556578947	0.734210526
Variance	0.014155702	0.01376462
Number of Observations	19	19
<i>t</i> Denominator	0.038333936	
Null hypothesis Mean Difference	0	
df	36	
<i>t</i> obtained	4.633794484	
<i>t</i> critical one-tail	1.6892	
<i>p</i> (<i>t</i> critical \leq <i>t</i> obtained) one-tail	0.0000229	
<i>t</i> critical two-tail	2.0294	
<i>p</i> (<i>t</i> critical \leq <i>t</i> obtained) two-tail	0.0000457	
Cohen's d effect size	1.503401453	

Output for a t test and Cohen's d using Excel

The next step is to run a statistical test. Each program uses different steps to perform the same test, with some requiring you to choose from various menu options (e.g., SPSS) and others requiring you to write a script or syntax (i.e., some lines of code, like a computer program). Excel uses

formulas that are beyond the scope of this book, but Neil Salkind has written an approachable primer for statistics using both Excel (Salkind, 2016) and R (Salkind & Shaw, 2019). When the analysis is complete, an output will show the results of the statistical procedure you just performed: You will need to learn how to interpret this output. [Figure 13.5](#) shows the output for a *t*-test using formulas in Excel. Note that the obtained *t* value is slightly different than the value we calculated earlier using means and standard deviations reported in the article—which had been rounded to two decimal places. Here, using the full dataset and carrying many decimal places, we can calculate *t* more precisely.



TRY IT OUT!

Look closely at the output for this analysis and try to find the following pieces of information:

1. What percentage of comprehension questions did the people who had a view of a person multitasking on their laptop get correct?
2. What about for those who couldn't see someone using their laptop?
3. What is the effect-size for this difference between groups?
4. How many people were in each condition?
5. What were the degrees of freedom and does this match what you would expect based on the total sample size and number of groups?
6. What is the *p*-value for a two-tailed test?
7. Is this *p*-value less than .05?

Selecting the Appropriate Statistical Test

How do you choose the appropriate statistical test for analyzing your data? There are many guides and tutorials you can access online, and your research supervisor should be able to help you with the decision. One important consideration is the scale properties of your variables (nominal, ordinal, interval, or ratio). In the tables below, we list which statistical tests are appropriate for which forms of data. We focus on variables that have (1) nominal scale properties, such as experimental and control conditions, or (2) continuous scores (i.e., interval/ratio scales) with many values, such as reaction time or rating scales. Some examples of variables with those scale properties are presented in *italics*.

Research Studying Two Variables

In these cases, the researcher is studying whether two variables are related, using non-experiments, quasi-experiments, or experimental designs. The Chi-square is a statistical test appropriate only when both variables are on a nominal scale ([Appendix B](#) provides details; note that “Chi” is pronounced similar to the “kai” in kite). The table below lists the appropriate tests for various combinations of independent and dependent variables.

IV (or Variable X)	DV (or Variable Y)	Statistical Test
Nominal <i>Buddhist/atheist</i>	Nominal <i>Vegetarian—yes/no</i>	Chi-square
Nominal (2 groups) <i>Procrastinators/non-procrastinators</i>	Interval/ratio <i>Academic average</i>	t-test

IV (or Variable X)	DV (or Variable Y)	Statistical Test
Nominal (3 groups) <i>Study strategy (re-reading, self-testing, and concept map conditions)</i>	Interval/ratio <i>Test score</i>	One-way analysis of variance
Interval/ratio <i>Optimism score</i>	Interval/ratio <i>Sick days last year</i>	Pearson correlation

Research with Multiple Independent or Predictor Variables

The following situations highlight some complex research designs. These designs have two or more independent or predictor variables that are studied with a single dependent or criterion variable. [Chapter 12](#) has a discussion of multiple regression.

IVs or Predictor Variables	DV or Criterion Variable	Statistical Test
Nominal (2 or more variables) <i>Cups of coffee (0, 1, and 2)</i>	Interval/ratio <i>Test score</i>	Analysis of variance (factorial design)
Interval/ratio (2 or more variables) <i>Optimism score/Average number of servings of vegetables eaten each week last year</i>	Interval/ratio <i>Sick days</i>	Multiple regression

There are many other types of designs and corresponding statistical analyses. Designs with multiple variables (that use multivariate statistics) are described in detail by Tabachnick and Fidell (2013). Consult Siegel and Castellan (1988) for procedures when using ordinal level measurement.

In addition to the type of response scale, statistical tests also have certain assumptions associated with them that must be met in order for it to be

appropriate to use the test. Many of these assumptions have to do with the shape of the data, whether it is drawn from a normally distributed population, for example. The t -test and F test both rely on the assumption that the data are drawn from a normal distribution. In contrast, the Chi-square test does not have this assumption, nor does regression. Thus, it is very important to perform your descriptive statistics and get a good sense for what your data looks like first, so you can choose the appropriate statistical test.



LO8 Integrating Descriptive and Inferential Statistics

The null hypothesis test is a pass/fail form of test: The p -value of the test statistic obtained is either below the alpha threshold set for statistical significance (and the null hypothesis is rejected), or above it (and the null hypothesis is retained). This hypothesis test tells us no information about whether an estimated effect is tiny or huge, or about the uncertainty around this estimate. For these judgments, effect-sizes and confidence intervals are needed (Cumming, 2014; Wilkinson et al., 1999). Some researchers are proposing that NHST be abandoned altogether in favour of these more nuanced values. As a result, effect-sizes and confidence intervals are becoming increasingly valued for the rich information they provide. Before we turn to a greater discussion of effect-sizes and confidence intervals, let us first review some of the critiques of NHST.

Although widely used, NHST and its associated p -values are widely misunderstood. Revisiting the definition of a p -value will help you to

understand the critiques of NHST. A *p*-value is defined as “[t]he probability that data like what we have observed (or more extreme data) could arise, if the null hypothesis were true” (Fraley & Marks, 2007). Note that the *p*-value tells us about the probability of observing data similar to what we have observed (i.e., the data collected and being analyzed), and does not tell us anything about the probability of the research hypothesis or the null hypothesis. In fact, it assumes that the null hypothesis is true. A small *p*-value, therefore, tells us that if the null hypothesis is true (and only if it is true), then data like what we have observed are unlikely. But what we really want to know is, “Now that I have observed these data, does the likelihood of the null hypothesis or research hypothesis change?” Although some believe you can transform the true definition of *p*-value into an inference regarding the probability of the null, this is based on faulty logic (Fraley & Marks, 2007). Thankfully, there is growing awareness that NHST cannot provide any information regarding the probability of the null or research hypothesis. Bayesian statistics, in contrast, can provide estimates of whether the null or research hypothesis are more or less likely in light of the data observed. For those interested in learning more about Bayesian statistics, there are several good introductory articles (e.g., Dienes, 2011; Etz & Vandekerckhove, 2018; Lindley, 1993), in addition to a free and easy-to-use software program for conducting Bayesian analyses, called JASP (<https://jasp-stats.org>).

Not only are *p*-values uninformative with respect to the probability of our hypotheses, they also do not tell us whether a result is likely to replicate or not (Cumming, 2008; Fraley & Marks, 2007). In fact, observing a *p*-value of .05 is a very poor indication of whether an exact replication of the same study would also produce a similar *p*-value. Goodman (2008) has published a succinct article outlining 12 misconceptions surrounding the interpretation of *p*-values, including the idea that a *p*-value above .05 indicates no difference between groups. As a result of the growing awareness of the problems with NHST, the American Statistical Association recently released a public statement on the issue, which includes many valuable references to relevant critiques (Wasserstein & Lazar, 2016). All of these problems with NHST have led some to suggest that *p*-values, and NHST, be abandoned completely (McShane et al., 2019; Trafimow & Marks, 2015). However, it remains the dominant practice

within many behavioural sciences, for the time being. At the very least, almost everyone seems to agree that including additional information, such as effect-sizes and confidence intervals, along with p -values is a welcome practice. In fact, including this information has been recommended by the American Psychological Association (APA) for the past two decades (Wilkinson et al., 1999).Page 271

Effect-Size

The concept of effect-size was discussed in [Chapter 12](#) as a way to describe the strength of the relationship between variables. In addition to reporting whether there was a statistically significant effect of the independent variable, the APA recommends that effect-sizes be reported and discussed in scientific reports (Wilkinson et al., 1999). Effect-sizes provide different information about an effect than the result of a statistical significance test. For example, in two-group experiments, the effect-size Cohen's d is often reported alongside the results of a t -test. There are many other types of effect-sizes: in fact, the Pearson correlation coefficient also functions as an effect-size, with larger correlations indicating a larger association or effect. Other effect-sizes operate like r^2 , indicating the amount of variance in one variable that can be explained by the variability in another variables (and vice versa). Different effect-sizes are appropriate for different types of research designs and statistical tests. [Appendix B](#) has the formulas for Cohen's d as well as omega squared, with the latter used to calculate effect-sizes in designs with three or more groups, including factorial designs.

It is possible for effects of any size to be statistically significant, when sample sizes are large enough. Because of how the p -value is calculated, any effect that is not *exactly* zero can become statistically significant with a large enough sample. For example, data from the World Values Survey show a statistically significant difference in life satisfaction between Australian and Mexican respondents. Australians ($n = 1465$) rated their life satisfaction ($M = 7.20$, $SD = 2.05$) lower than Mexican respondents ($n = 2000$, $M = 8.51$, $SD = 1.93$), $t(3463) = 19.19$, $p < .0001$. For the sake of example, let's assume we had only one-tenth the number of participants (15 Australians and 20 Mexicans), but the same means and standard deviations (SD). By simply changing the sample size, and nothing

else, the result becomes no longer statistically significant, $t(33) = 1.92$, $p > .05$. This demonstrates how statistical significance is impacted by sample size. In contrast, effect-size estimates are not affected by sample size. In both cases, our estimate of the effect-size is a Cohen's $d = .66$. This makes it obvious that p -values and effect-sizes are not the same thing, but in fact give different kinds of information.

In addition, some statistically significant differences might have very little *practical* significance, or not be very meaningful. For example, if an expensive new psychiatric treatment reduced the average hospital stay from 60 days to 59 days, it might not be practical to use this costly technique even if the difference is statistically significant. In this example, an additional day of hospitalization would likely cost less than the proposed treatment. Not only does statistical significance have no relation to practical or meaningful significance, effect-size is also unrelated to practical significance. Sometimes a very large effect can be unimportant, or a very small effect could be incredibly important. Let us consider an example in which a treatment with a very small effect-size has considerable practical significance. This could occur when a very large population is affected by a fairly inexpensive treatment. Suppose a simple flextime policy for employees reduces employee turnover by 1 percent per year. This doesn't sound like a large effect. However, if a company normally has a turnover of 2,000 employees each year, and the cost of training a new employee is \$10,000, the company saves \$200,000 per year by allowing flextime. This amount of money may have an important practical significance for the company. There is no statistic that tells you whether an effect is important, meaningful, or of practical significance. Instead, we must rely on reason to decide whether this is the case, carefully considering the context and the outcome. When outcomes are serious, even small effects can be important. If a large effect doesn't have much influence on any important outcome for this context, it is probably not very important. As a result of the relations between statistical significance, effect-size, and practical or meaningful significance, it is insufficient to consider the result of a null hypothesis test alone. It is important to also consider effect-size and, perhaps, consider the context and outcomes of any result.

Page 272

Confidence Intervals and Statistical Significance

After obtaining a sample value (e.g., a mean or an effect-size), we can calculate a confidence interval to define the most likely range of actual population values (Cumming, 2012; Kline, 2013). As discussed in [Chapter 7](#), the true definition of a confidence interval is complicated and frequently misunderstood (for the true definition and a discussion, see Hoekstra et al., 2014), but a 95 percent confidence interval behaves in practice as an 83 percent replication interval (i.e., a replication value will fall within the 95 percent confidence interval, 83 percent of the time; Cumming, 2008). A 99 percent confidence interval will provide greater certainty, but the range of values would be larger (i.e., the interval will be wider).

For example, a confidence interval can be obtained for each of the means in the laptop multitasking experiment (Sana et al., 2013). Although confidence intervals can be calculated in different ways and the exact calculations are beyond the scope of this book, it is worth noting that bootstrapped confidence intervals are typically preferred. For now, just consider the 95 percent and 99 percent confidence intervals around the means of the two conditions:

	Viewing Multitasking Laptop Group	Not Viewing Multitasking Laptop Group
Obtained sample value (mean proportion correct)	.56	.73
95% confidence interval around the mean (lower bound, upper bound)	(.50, .61)	(.68, .79)
99% confidence interval around the mean (lower bound, upper bound)	(.48, .64)	(.66, .81)

First, you might notice that the confidence intervals for the two means do not overlap, which is a clue that the difference is statistically significant. Examining confidence intervals can be an alternative way of thinking about statistical significance.

However, examining these confidence intervals provides you with far more information, beyond statistical significance. Although the obtained sample means are your best estimate of the population means, these confidence intervals provide you with information about the uncertainty around your estimate. The size of a confidence interval (i.e., how wide it is, indicating greater uncertainty) is related to both the size of the sample and the confidence level (e.g., 95 percent or 99 percent confidence). As sample size increases, the confidence interval narrows because the population value is being estimated more precisely.

Assuming the sample size stays the same, notice how the interval widens—indicating less precision and more uncertainty in our estimate—as the confidence level increases from 95 percent to 99 percent. According to APA standards, confidence intervals should be reported for each point estimate (i.e., an estimate of a single value, like a mean or an effect-size), whenever possible (Wilkinson et al., 1999). This is because these confidence intervals provide a richer, more informative, alternative to the standard pass/fail null hypothesis approach (Cumming, 2012; Kline, 2013; Masson & Loftus, 2003). Confidence intervals, for example, tell us something about the likelihood of a particular result replicating if the study were to be repeated exactly, multiple times (Cumming, 2008). In contrast, *p*-values do not tell us anything about whether a particular effect will replicate, by definition (Fraley & Marks, 2007), nor do they do well at this prediction in practice (Cumming, 2008).Page 273

Conclusion Validity

Conclusion validity is the extent to which the conclusions about the relationships among variables reached on the basis of the data are correct or reasonable (Trochim, 2006). Conclusion validity is a requirement for conclusions drawn from quantitative data as well as from qualitative data ([Chapter 6](#)). When working with quantitative data, interpreting the effect-size and confidence intervals along with significance test results and study design details, will help us to draw valid conclusions from our data. In closing this section, a quote from the American Statistical Association's statement on *p*-values seems very apt:

Good statistical practice, as an essential component of good scientific practice, emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean. *No single index should substitute for scientific reasoning* (Wasserstein & Lazar, 2016; emphasis added).

The Importance of Replications

Throughout this discussion of statistical analysis, we have focused mainly on the results of a single study. What were the means and standard deviations? Was the mean difference between groups statistically significant? How narrow or wide are the confidence intervals around the estimates of the population means and the effect-size estimates? However, scientists do not attach too much importance to the results of a single study. A rich understanding of any phenomenon comes from the results of numerous studies investigating the same variables. Instead of inferring population values on the basis of a single investigation, we should look to the results of several studies that are able to replicate a particular finding (Cohen, 1994). The importance of replications is a central concept in [Chapter 14](#). In that final chapter, we will examine various issues of generalizing research findings beyond the specific circumstances in which the research was conducted.

Study Terms

Test yourself! Define and generate an example of each of these key terms.

- [alpha level](#) (p. 257)
- [conclusion validity](#) (p. 273)
- [error variance](#) (p. 259).
- [F test](#) (p. 259).
- [inferential statistics](#) (p. 251)
- [null hypothesis](#) (p. 253)
- [power](#) (p. 265)
- [probability](#) (p. 253)
- [publication bias](#) (p. 263)
- [research hypothesis](#) (p. 253)
- [sampling distribution](#) (p. 255)
- [statistically significant](#) (p. 252)
- [systematic variance](#) (p. 259)
- [t-test](#) (p. 257)
- [Type I error](#) (p. 260)
- [Type II error](#) (p. 261)

Review Questions

Test yourself on this chapter's learning objectives. Can you answer each of these questions?

1. Distinguish between the null hypothesis and the research hypothesis. When does the researcher decide to reject the null hypothesis? How are sampling distributions involved in this decision?
2. What is meant by statistical significance? What elements influence whether obtained results will be significant?
3. List and describe the three basic steps involved in using the *t*-test.
4. Compare and contrast the *t*-test and the *F* test. Consider their numerators and denominators, and reasons to choose either test.
5. Distinguish between Type I and Type II errors. How does alpha level relate to the probability of making a Type I error if the null hypothesis is true?
6. How do Type I and Type II errors influence the accuracy of our published research overall?
7. What influences the probability of a Type II error?
8. What is the difference between statistical significance and practical significance? Between statistical significance and effect-size?
9. How does effect-size relate to practical or meaningful significance?
10. Discuss reasons why a study might show non-significant results.

11. Compare information obtained using a null hypothesis test (e.g., a *t*-test) with information from effect-sizes and confidence intervals. What information does each analysis provide?

Deepen Your Understanding

Develop your mastery of these concepts by considering these application questions. Compare your responses with those from other people in your study group.

1. In an experiment, one group of research participants is given ten pages of material to proofread for errors. Another group proofreads the same material on a computer screen. The dependent variable is the number of errors detected in a five-minute period. A .05 alpha level is used to evaluate the results.
 1. What statistical test would you use? Why?
 2. What is the null hypothesis? The research hypothesis?
 3. What is the Type I error? The Type II error? Describe your answer in words.
 4. What is the probability of making a Type I error if the null hypothesis is false?
2. Professor Anunoby collected data using the design in Question 1. The average number of errors detected in the Print and Computer conditions was 38.4 and 13.2, respectively; this difference was not statistically significant. When Professor Siakam conducted the same experiment, the means of the two groups were 21.1 and 14.7, but the difference was statistically significant. Explain how two researchers using the same method could arrive at these different conclusions about the null hypothesis. Page 275
3. Suppose that you work for the child social services agency in your province (e.g., Children's Aid Foundation). Your job is to investigate instances of possible child neglect or abuse. After collecting evidence from a variety of sources, you must decide whether to leave the child in the home or place the child in protective custody. Specify the null

and research hypotheses in this situation. Describe in words what the Type I and Type II errors are. Is a Type I or Type II error more serious in this situation? Why?

4. A researcher investigated attitudes toward people in wheelchairs. Would people react differently to a person they perceived to be confined temporarily to the wheelchair compared to a person who had a permanent disability? Participants were randomly assigned to two groups. In one group, people worked on various tasks with a confederate in a wheelchair. In the other group, people worked with the same confederate in a wheelchair, but the confederate wore a leg cast. After the session was over, participants completed a questionnaire regarding their reactions to the study. One question asked, “Would you be willing to work with your test partner in the future on a class assignment?” with “yes” and “no” as the only response alternatives. What would be the appropriate significance test for this experiment? Why? Recalling our discussion from [Chapter 9](#) regarding sensitivity, can you offer a critique of the dependent variable? If you changed the dependent variable, would it affect your choice of significance tests? If so, how?

Generalizing Results



©Todd Ryburn Photography/Getty Images

Our first instinct when we learn new information is to take it and run with it, but there are some obstacles to generalizing what we learn from specific samples. Fret not, in this chapter we'll go over these obstacles, some ways to overcome them, and ultimately introduce ways you can take your newfound knowledge of research methods and soar, like a regal Canada goose!

Learning Objectives

Keep these learning objectives in mind as you read to help you identify the most critical information in this chapter.

By the end of this chapter, you should be able to:

1. LO1 Discuss four challenges to generalizing research results to populations other than the population sampled.
2. LO2 Identify three specific aspects of a study's procedure that may affect the ability to generalize beyond that study, and suggest possible solutions.
3. LO3 Discuss the importance of replications, distinguishing the procedures and uses of direct replications and conceptual replications.
4. LO4 Evaluate the strengths and limitations of using convenience samples, and identify ways to find more diverse samples.
5. LO5 Distinguish between narrative literature reviews and meta-analyses.
6. LO6 Identify ways to continue using your knowledge of research methods.

Page 277 What do research results tell us about populations? Our ability to generalize research findings has been a source of heated debate through the years. A single study is conducted with a particular sample and in a particular context. The *external validity* of that study is the extent to which the results can be generalized beyond that study and sample to other populations and settings. The issue of external validity is complex, raising deep questions about the usefulness of the knowledge we create when doing research. Can results be generalized to other populations or to other study contexts? In this chapter, we will explore some ways to tackle these questions, and close by considering ways to generalize your knowledge beyond this book.

Challenges to Generalizing Results

In this section, we consider two ways in which research results may or may not generalize beyond the sample studied. Can these results generalize beyond your sample of participants to other populations? Can the results generalize beyond the particular situation of your study to other settings and situations?



LO1 Can Results Generalize to Other Populations?

Except in rare cases, research is conducted on samples rather than entire populations. If we wish to generalize a set of results to all of humanity (i.e., the entire population of humans on earth), the best—but completely impractical—method for doing so would be to randomly select a sample from the earth’s entire population. In contrast, most psychological research relies on *convenience samples*, people selected simply because they are available and willing ([Chapters 7](#) and [9](#)). Not surprisingly, the population most available to university researchers is university students. University students are bound to be different from the general population in a country in many ways. Do research findings based on these students tell us only about other, similar students, or do they also tell us about the general

population in that country, or even all of humanity? In other words, can we generalize our findings to a wider population?

Beyond University Students

A substantial portion of research in psychology is based upon undergraduate students. Estimates based on reviews of scientific journals suggest that about 68 percent of studies use university undergraduates as participants (Sears, 1986; Smart, 1966; Wintre, North, & Sugar, 2001). The problem is that university students represent a highly restricted population: typically, first- and second-year students taking an introductory psychology class (Sears, 1986; Wintre et al., 2001). They therefore tend to be young and to possess the characteristics of late adolescence: a developing sense of self-identity, social and political attitudes that are in a state of flux, a high need for peer approval, and unstable peer relationships. They are also intelligent, have good cognitive skills, and know how to win approval from authority (having done well enough in a school environment to get into university). What we know about “people in general” may actually be limited to a highly select and unusual group, both culturally and historically. Researchers at the University of British Columbia have described this group as WEIRD: Western, Educated, Industrialized (as in, from an industrialized country), Rich, and Democratic (Henrich, Heine, & Norenzayan, 2010a). As a result of being WEIRD, the conclusions drawn from these participants seem unlikely to generalize to all of humanity.

Beyond Volunteers

Researchers usually must ask people to volunteer to participate in their research. At many universities, introductory psychology students are often required to volunteer for experiments for course credit. If you are studying populations other than university students, you are even more dependent on volunteers. Recruitment for these populations might involve asking parents visiting a science centre to participate in a study on parenting, or asking users of a particular Internet forum to complete a survey online. Research on volunteers has found that those who agree to participate often differ in various ways from non-volunteers (Rosenthal & Rosnow, 1975). Volunteers tend to be more highly educated, more in need of approval, and more social

(Rosenthal & Rosnow, 1975). They also tend to have higher levels of trait conscientiousness and lower levels of trait neuroticism than non-volunteers (Lönnqvist et al., 2007).Page 278

Further, different kinds of people volunteer for different kinds of studies. In most universities, there is a sign-up board for potential studies or a website with different studies listed: Both will include key words indicating what the study is about. Different people may be drawn to different types of studies, with some finding a study on “problem solving” more attractive than one on “interaction in small groups,” whereas others might have the opposite preference. These study titles do seem to influence who signs up (Hood & Back, 1971; Silverman & Margulis, 1973). Another factor that can influence who signs up for a study is whether some financial compensation is offered. Studies that recruit participants by emphasizing financial rewards tend to attract less-altruistic participants than studies emphasizing the potential learning opportunity (Krawczyk, 2011). Similarly, studies offering course credit for participating tend to attract less-motivated participants than studies that do not offer course credit (Sharp, Pelletier, & Lévesque, 2006), although offering either money or course credit boosts volunteer rates more than offering nothing. There are many factors related to how a study is advertised that influences what kinds of volunteers are likely to sign up.

Beyond the Gender of Participants

Sometimes, researchers use either mostly females or males, or only one gender, simply because this is convenient or the procedures seem better suited to one gender. Similarly, individuals who identify as a gender minority (e.g., as transgender or genderqueer) might be excluded from analysis. Given the possible differences between the genders, however, the results of such studies may not generalize equally to all genders (Denmark, Russo, Frieze, & Sechzer, 1988). For example, in past research, research on male rats was found not to generalize to female rats; in fact, female rats exhibited the exact opposite pattern (Barha, Paluwski, & Galea, 2007).

In psychological research with humans, gender could influence results and subsequently the conclusions drawn from these results. Denmark and

colleagues (1988) have identified several ways that gender bias may arise throughout the research process. If a sample is solely or predominantly one gender, it is best to avoid concluding that a widely generalizable truth has been found. It is possible that the variables relate differently, or that people with different gender identities interpret experimental materials differently. If you wish to generalize the results of a study across many different genders, it is necessary to replicate the study with participants of the gender in question. It is also possible, when a sample includes a good balance of genders, to include gender as a variable in the analyses to investigate if the results generalize across the genders studied.

Beyond Culture

Not so long ago, participants in most studies were North American university students who were primarily White, reflecting the predominant population of university students. Today, however, many samples of university students are ethnically diverse because the population of university students across North America has become increasingly diverse. In addition, more and more psychological research is being done in countries around the world. As a result, the overall external validity of research has improved. It is also now much easier to compare ethnic groups, to examine cross-cultural differences and similarities. In the late 1980s, fewer than 10 percent of studies in social psychology included comparisons of two or more cultures (Pepitone & Triandis, 1987). Since then, there has been an explosion of interest in studying different cultures. Miller (1999) has encouraged psychologists to take a broad view of the importance of culture, in which “culture is understood as shared meaning systems that are embodied in artifacts and practices and that form a medium for human development” (p. 86). Even more recently, a special section in the journal *Perspectives on Psychological Science* was devoted to highlighting “the fundamental nature of culture research in psychological science,” including how researchers are revising and developing new theories for a better, more global, and more relevant science (Gelfand & Diener, 2010, p. 390). Despite this progress in comparing psychological phenomena across cultures, psychology remains grounded in a Western perspective. A truly global psychology, arguably, will require greater

incorporation of concepts and methods originating in non-Western cultures (Berry, 2013).Page 279

So far, much of cultural research has centred on identifying cross-cultural similarities and differences in responses to the same environments, along with personality and other characteristics (Matsumoto, 1994). For example, Buunk and colleagues (2010) conducted a series of studies to examine the influence of parents on dating and marriage across cultures. They asked students in the Netherlands, Iraq, and Canada to rate the degree to which their parents controlled their choice of romantic partner. Because the Netherlands is highly individualistic and Iraq is highly collectivistic, researchers hypothesized that parents would have far more say in marriage partners in Iraq than in the Netherlands. Their hypothesis was confirmed, and the study uncovered a very large effect-size (Cohen's $d = 2.23$; [Chapter 12](#)). In their Canadian sample, Buunk and colleagues compared University of British Columbia students whose cultural background was East Asian (i.e., more collectivistic) with those whose background was European (i.e., more individualistic). Once again, students with a collectivist cultural background reported more parental involvement in their choice of romantic partner than students with an individualist cultural background (Cohen's $d = 1.42$). This type of research informs us about the generality of effects across cultural groups.

☆ Student Spotlight: Generalizing across Cultures ☆

Distinct cultures can exist even within a single country, and it can be fruitful to investigate whether the effects observed in one culture generalize to another. Taryn Buoy, working with Dr. Elena Nicoladis at the University of Alberta, was interested in code-switching: when bilingual individuals switch from one language to another within the same conversation. More specifically, they were curious as to whether French-English bilinguals from Quebec might be more critical of code-switching than those from Alberta, as past research has found that the Quebecois might value purity when it comes to language. To research this question, they asked bilinguals from both Alberta and Quebec to view and evaluate short video clips in which people engaged in code-switching or maintained their use of a single

language. You can read more about their research in the *Journal of Intercultural Communication Research* (Buoy & Nicoladis, 2018).

Operational definitions of the constructs we study are grounded in particular cultural meanings (Byrne & van de Vijver, 2010). A measure of self-esteem that is appropriate for an individualistic culture is probably not appropriate for use, and would yield misleading results, in a collectivistic culture (Hamamura, Heine, & Paulhus, 2008; Heine, Lehman, Peng, & Greenholtz, 2002). It is therefore crucial to reconsider construct validity ([Chapter 5](#)) whenever using a familiar operational definition in a new population.

Page 280



LO2 Can Results Generalize beyond the Specific Study Situation?

Beyond the Experimenter

The person who actually conducts the experiment can trigger another generalization problem. In some studies, only one experimenter is used to reduce variability in how the experimenter influences participants. Because little attention is typically paid to the personal characteristics of experimenters (McGuigan, 1963; Strohmetz, 2008), it is possible that the results of a study using only one experimenter cannot be generalized to other types of experimenters.

Some of the important characteristics of experimenters include personality, gender, and amount of practice in the role of an experimenter (Kintz, Delprato, Mettee, Persons, & Schappe, 1965). A warm, friendly experimenter may produce different results than a cold, unfriendly experimenter. Participants are also more productive and cooperative when experimenters are dressed in accordance with stereotyped gender roles (Green, Sandall, & Phelps, 2005). It has even been shown that rabbits learn faster when trained by experienced experimenters (Brogden, 1962)!

One solution to the problem of generalizing to other experimenters is to use two or more experimenters, with differing characteristics (Rubin, 1975). Another option is to deliver instructions using a computer, which minimizes the amount of interaction between experimenters and participants, thereby reducing the potential for influence (Strohmetz, 2008; [Chapter 9](#)).

Beyond a Pretest

Researchers must often decide whether to give a pretest ([Chapter 8](#)). Intuitively, pretesting seems to be a good idea. The researcher can examine whether groups are equivalent on the pretest, and sometimes it is important to examine changes in people's scores from pretest to posttest, rather than simply comparing posttest scores. In longitudinal studies that have the risk of participants withdrawing from the study, a pretest allows us to look for any effects of selective attrition. A pretest lets us determine whether the people who withdrew from the study were different from those who completed it.

Pretesting, however, may limit the ability to generalize any results to populations that do not receive a pretest. In the real world, people are rarely given a pretest. For example, people do not regularly take stock of their attitudes before listening to a political speech or viewing an advertisement (cf., Lana, 1969). One specific research design, the Solomon four-group design (Solomon, 1949), can be used when a pretest is desirable, but there is concern that this pretest will affect later responses. In the Solomon four-group design, the same experiment is conducted with and without the pretest. The researcher can then examine whether there is an interaction

between the independent variable and the pretest variable. If the pretest has no effect, posttest scores on the dependent variable are the same regardless of whether or not the pretest was given.

Beyond the Laboratory

Research conducted in a laboratory setting has the advantage of allowing the experimenter to study variables under highly controlled conditions. In experiments, the goal of high internal validity (allowing for the ability to infer causality) may sometimes conflict with the goal of external validity. Does the artificiality of the laboratory setting limit the ability to generalize results to real-life settings? Field experiments are one way that researchers try to examine phenomena under more realistic circumstances, and thereby increase the external validity of their experiments ([Chapter 4](#)). In a field experiment, the researcher manipulates the independent variable in some natural setting, like a factory, a school, or a cafeteria; the influence of this manipulation on some outcome is then measured.Page 281

Conducting research in both laboratory and field settings provides the greatest opportunity for advancing our understanding. Consider research on eyewitness testimony. Many important limits of the accuracy of eyewitness memory to a crime have been illuminated in the lab. For example, in one study, Carleton University students who viewed a videotaped crime were much less likely to correctly identify the perpetrator in a lineup if the perpetrator had changed his hair colour and style (Pozzulo & Marciniak, 2006). This finding is consistent with field research showing that perpetrator hair colour is among the most accurately remembered details for eyewitnesses, even in violent crimes (Wagstaff et al., 2003). Yet controversy remains over whether lab-based findings can be generalized to real events in which people witness actual crimes live (Yuille, Ternes, & Cooper, 2010), with the latter being a far more stressful situation (Pozzulo, Crescini, & Panton, 2008). Both lab- and field-based studies are vital contributors to our understanding of eyewitness testimony.

Do laboratory and field experiments that examine the same variables generally produce the same results? To answer this question, Mitchell (2012) found 82 meta-analyses of lab studies and field studies that both

examined the same topic, allowing for 217 separate comparisons. This allowed them to investigate whether the lab and field studies found the same result, and whether the two types of studies differed in their estimation of the effect-size for the phenomenon ([Chapter 12](#)). These meta-analyses reflected topics from a variety of subfields in psychology, mostly from social and industrial/organizational psychology, but also from personality psychology, psychometrics, and other fields. Overall, laboratory and field experiments found similar results: The correlation between the effect-size found in the lab and the field was high ($r = .71$). However, for 30 of these comparisons (based on meta-analyses), the lab and field studies found completely opposite results. The correspondence between lab and field studies for the two groups with the most data was $r = .89$ for industrial/organizational psychology, and $r = .53$ for social psychology. When the effect-sizes being reported were small, as is the case for studies of gender differences, the correspondence between lab and field studies was poor. In summary, although lab studies and field studies do tend to find similar effects, there is substantial variability in the degree to which this is true, based on the field of psychology and topic of study. As we discuss next, when findings are replicated in multiple settings, our confidence in the generalizability of the findings increases.

Solutions to Generalizing Results



LO3 Replicate the Study

[Replication](#) is an important way to overcome some of the questions of generalization that stem from the results of a single study. There are two types of replications to consider: direct replications and conceptual replications (Schmidt, 2009).

Replicate Directly

A [direct replication](#) is an attempt to replicate the procedures of a study as closely as possible to see whether the same results are obtained. In best practice, individual researchers attempt to directly replicate their own work when possible, especially when the results from the initial study are unexpected or are based on small samples (Cesario, 2014). Direct replications are crucial for determining whether an original finding can generalize to other samples drawn from the same population (e.g., undergraduates at an institution)—in other words, to offer evidence that the initial result was not simply a Type I error (concluding there is an effect, when no effect is present in reality; [Chapter 13](#); Simons, 2014).Page 282

Researchers may also engage in a direct replication when embarking on a new line of research. If you are just beginning your own research on a topic, you may try to replicate a crucial study to make sure that you understand the procedures

and can obtain the same results. Often, researchers attempt a direct replication and then pair this with a novel study that builds on that finding, known as a replication and extension (Bonnet, 2012). For example, a researcher might replicate a particular finding, then see if it can also be observed in a different culture or age group. When you can replicate an original research finding using the same or very similar procedures, confidence in the generality of the original findings is increased.

However, at times, a researcher will be unable to replicate a previous finding. A single failure to replicate does not always mean that the original phenomenon does not truly exist, much as a single study demonstrating an effect should not convince us on its own that a phenomenon is real. Failures to replicate share the same difficulties of interpretation as statistically non-significant results, discussed in [Chapter 13](#). A failure to replicate could mean that the original results are invalid, but it could also mean that the replication attempt was flawed. For example, if the replication is based on the procedure as reported in a journal article, it is possible that the article omitted an important aspect of the procedure. For this reason, it is usually a good idea to contact the researcher to obtain detailed information on all materials and procedures used in the study. However, just as a set of well-designed studies using large samples can provide evidence in favour of a null, a well-designed replication attempt with a large sample that fails to detect an effect should also be considered evidence that the effect may not truly exist. In some cases, the effect may only occur under certain, limited circumstances. In other cases, it may turn out that the original researcher misrepresented the results in some way, or engaged in questionable research practices in order to produce false evidence for a phenomenon (for a discussion of scientific fraud, refer to [Chapter 3](#)).

The so-called “Mozart effect” offers an example of the importance of replications. In the original study, university students listened to ten minutes of a particular Mozart sonata and subsequently exhibited better performance on a spatial-reasoning task compared to controls (Rauscher, Shaw, & Ky, 1993). Two years later, the same team reported a replication of the effect using a different measure of spatial ability (Rauscher, Shaw, & Ky, 1995). Despite the fact that the effect was temporary, lasting about ten minutes, these findings received a great deal of attention in the press. People quickly generalized this finding to conclude that one could increase a child’s intelligence by playing Mozart sonatas. In fact, one U.S. governor began producing Mozart CDs to distribute in maternity wards, and entrepreneurs began selling Mozart kits to parents. Over the next few years, however, there were many failures to replicate the Mozart effect (see Steele,

Bass, & Crook, 1999). Rauscher and Shaw (1998) responded to the many replication failures by precisely describing the conditions necessary to produce the Mozart effect. However, other researchers remained unable to obtain the effect even though they followed these detailed procedures (McCutcheon, 2000; Steele et al., 1999). A meta-analysis aggregating the results of 40 studies, based on around 3,000 participants, concluded that there was a small effect of listening to Mozart on spatial reasoning compared to not listening to music (Pietschnig, Voracek, & Formann, 2010). However, listening to any music at all was found to result in an effect of equal size, likely due to music increasing arousal levels (e.g., Thompson, Schellenberg, & Husain, 2001). It seems there is nothing special about listening to Mozart after all, but simply an effect of arousal on task performance. Page 283

☆ Student Spotlight: Direct Replications ☆

Research on attachment avoidance typically finds that individuals high in this trait prefer to avoid intimacy. However, some studies have found that those high in avoidant attachment may respond positively to intimacy, under certain circumstances. Aviva Philipp-Muller, an undergraduate at the University of Toronto, wanted to try to replicate one such finding (MacDonald & Borsook, 2010), working closely with one of the researchers who produced this original finding, Dr. Geoff MacDonald. They followed the methods of the original study very closely, but added some additional measures after the replication part of the study was completed. In this way, this study can be viewed as a replication and an extension. To find out if they were able to replicate the results of the original study, you can read their report of this direct replication attempt in the *Journal of Experimental Social Psychology* (Philipp-Muller & MacDonald, 2017).

Direct Replication and Disciplinary Reform

Several high-profile cases of failures to directly replicate research results over the last few years have shone a spotlight on the issue of replication. In one case, a paper published in a major journal purportedly offered evidence of extra-sensory perception (Bem, 2011). In these studies, participants were better than chance at guessing the on-screen location of hidden pictures, but only when the pictures were arousing: Their future emotions seemed to affect choices in the present. When skeptical researchers attempted to replicate these findings, many were unable to do so (for a summary, see Baruss & Rabier, 2014). As another example, several attempts failed to replicate very well-known studies of non-conscious

priming. In one of the original studies, people who were primed with the concept of old age subsequently walked more slowly, suggesting that a subtle prime had a measurable effect on behaviour (Bargh, Chen, & Burrows, 1996). One group of researchers did manage to replicate the effect, but only when the experimenters believed that it would occur (Doyen, Klein, Pichon, & Cleeremans, 2012). This suggests that the original result may have resulted from an experimenter expectancy effect ([Chapter 9](#); for other failed replications of similar studies, see Harris, Coburn, Roher, & Pashler, 2013; Pashler, Coburn, & Harris, 2012; see Yong, 2012, for an overview of the controversy).

These failures to replicate precognition and behavioural priming effects were some of the major triggers of recent debate and reform efforts (see also [Chapter 3](#) for cases of fraud). Many complex questions have been raised about the way research in psychology—and other disciplines—is conducted. Consider the following questions about replication in particular: Whose responsibility is it to replicate results? To what extent can we trust past literature that no one has attempted to replicate? What is the best way to replicate others' work? Which kinds of replications are most important? Answers to these questions will continue to emerge and change in the coming years; here we consider some recent attempts.

Historically, direct replications have been difficult to publish, particularly in the social sciences (Fanelli, 2012). Recent commitments made by journal editors in psychology are changing that norm by accepting particular types of replication studies after adequate peer review. Researchers are also developing a set of best practices that are helping to ensure that direct replications are high quality and objective (see Brandt et al., 2014). Some recommendations for convincing replication attempts include ensuring high statistical power (by employing a very large sample size; [Chapter 13](#)); following the original procedures as closely as possible, including using the original materials if obtainable; and making the details public for other researchers to verify (including the original study's authors). For example, Earp and colleagues (2014) attempted to directly replicate a widely cited study examining morality and physical purity. They used the original materials and ensured that their samples were ethnically diverse and large enough to detect any effects that might exist. Yet they failed in three studies to find any evidence of the effect. In the past, studies that fail to find an effect would typically end up in the file drawer. But now, in light of growing awareness of the importance of direct replications, they were able to successfully publish their methodologically rigorous demonstrations of a null result.

Page 284

It is becoming increasingly common for researchers from many labs to collaborate, in order to conduct simultaneous direct replications. Earp and colleagues (2014) attempted replications in the United States, the United Kingdom, and India. Not only does a multi-labs approach add external validity to the results, it also promotes objectivity as researchers are less likely to have a personal investment in the outcomes. The journal *Perspectives on Psychological Science* has created a special type of paper called a Registered Replication Report (RRR), which requires the cooperation of many labs to directly replicate important findings, as well as cooperation from the target study's original author whenever possible to approve the method and supply original materials (Simons, Holcombe, & Spellman, 2014). Importantly, the study design is pre-registered with the Open Science Framework (see <https://osf.io/>), the methodological protocol undergoes peer review before data are collected, and the result is virtually guaranteed publication regardless of whether the original result is supported or not. The first RRR was published in September 2014 (Alogna et al., 2014), it involved over 30 labs from 10 countries, and it ultimately replicated the original results, with some qualifications (for a commentary from the original study's first author, see Schooler, 2014). The RRR and other published direct replications represent a major commitment to high-quality psychological science. In addition to many labs collaborating to replicate a single effect, there have also been several large-scale attempts to replicate many different findings in psychology (e.g., 100 different results), also by forming coalitions across many different labs (e.g., Camerer et al., 2018; Klein et al., 2018; Open Science Collaboration, 2015). These valuable attempts to characterize the reproducibility of behavioural science results help shine a light on what findings appear to be robust, and which may require reconsideration or further research. Moreover, these collaborations demonstrate that many scientists can band together to achieve remarkable things, things that could never be achieved by one or two labs working on their own.

Replicate Conceptually

In contrast to a direct replication, which strives to replicate the procedures of a past study as closely as possible, the use of different procedures to replicate a research finding is called a *conceptual replication*. In most research, the goal is to discover whether a relationship between conceptual variables exists. The music manipulation in the original Mozart study mentioned earlier used the first section of Mozart's *Sonata for Two Pianos in D Major* (K. 448). This particular selection of music is a specific operationalization of the independent variable. Likewise,

the specific task chosen as the dependent measure is an operationalization of a more general concept: spatial reasoning.

In a conceptual replication, the independent variable is manipulated in a different way and/or the dependent variable is measured in a different way from the original study. A relationship that appears with one set of operationalizations should generalize to different ways of manipulating and measuring the same variables. Sometimes a conceptual replication may involve an alternative stimulus (e.g., a different Mozart sonata, or music by a different composer) or an alternative dependent measure (e.g., a different spatial-reasoning task). When conceptual replications produce similar results as the original study, a case can be made that the relationship between the variables generalizes beyond the original operationalizations. Page 285

This discussion should also alert you to an important way of thinking about research findings. The findings represent relationships between conceptual variables but are grounded in specific operationalizations. You may read about the specific methods employed in a study and speculate whether they are so unusual that they could never generalize to other situations, or to other operationalizations. A conceptual replication, however, would demonstrate that the relationship between the conceptual, theoretical variables is still present.

Conceptual replications can help develop theories of behaviour; however, they are not a substitute for direct replications (Schmidt, 2009). One problem with conceptual replications is the potential to promote Type I errors. If a conceptual replication fails to find an effect, it is possible to toss that study aside as methodologically problematic, and continue trying different operationalizations until the original effect is conceptually replicated (LeBel & Peters, 2011). This is one of the major criticisms of the extra-sensory perception paper (Bem, 2011). Despite offering nine conceptual replications, there were also many “pilot studies” that had failed to show effects and were dismissed as methodologically problematic, simply by virtue of them failing to show the effect. By deciding whether a methodology is sound based on whether the outcome supports your own hypothesis, one greatly increases the chance of Type I errors (i.e., concluding that an effect exists, one none actually exists in truth). Thus, for greater confidence in the generalizability of an effect, consider conducting a direct replication first, to ensure that the original results generalize beyond the original sample. Then, carefully substitute alternative operationalizations in a conceptual replication to develop a theory about how the underlying theoretical variables relate.

Consider Different Populations



LO4 Recognize the Limits of Convenience Samples and Seek Diverse Samples

Psychologists rely heavily on undergraduate students for their research participants. As we noted [earlier](#), this unique population does not represent the general population of the country from which the sample was drawn, nor does it represent humanity more generally (i.e., it is WEIRD; Henrich et al., 2010a, 2010b). This represents a challenge to the external validity of behavioural science. In [Chapter 7](#), we emphasized the importance of randomly sampling from a population when you wish to generalize to that population (e.g., all Canadians). If you wish to generalize your results to all of humanity, it would be ideal to randomly sample participants from the earth's population, but this is clearly impractical (and likely impossible).

Does that mean we should throw out all data that relies on undergraduate samples, because it is useless? Not at all! But we can question whether our results would generalize beyond this specific population, and seek broader samples whenever possible. Before dismissing research that uses any particular type of participant, such as university students, consider the characteristics of the sample under question. How well does it represent, demographically, the characteristics of your region or country? To the extent that a sample of undergraduates is diverse on a particular dimension (e.g., ethnicity), making generalizations to populations that share these dimensions is less problematic.

We should think about how any results might be similar or different, if a different population were investigated. University students, after all, *are* human—we just need to be mindful about how far we extend our claims. How might older people, younger people, or people from different cultural or socio-economic backgrounds respond to the study’s variables? Rather than simply assuming that processes operate universally, or not thinking about it at all, actively consider how different populations might think or act differently. Then use factorial designs to test those hypotheses and adapt the theory accordingly. Although we all have a shared humanity, there is enormous diversity in how we think and act, based on many individual differences (e.g., age, culture, environmental upbringing). One quite dramatic example of this idea comes from members of a small nomadic culture who live near Burma (Henrich et al., 2010a). These people have the ability to see clearly underwater because their pupils constrict underwater (rather than dilate, as do most peoples’ pupils). A thorough theoretical understanding of vision, that generalizes across all humanity, must therefore be able to account for both types of pupillary reactions. If we see complex variability for pupillary actions, thought to be basic biological process common across all humans, just consider the variability for psychological processes!Page 286

However, it is not always possible to collect data from diverse sources. That said, researchers should strive to do so whenever possible. Consider research conducted by Simon Fraser University’s Lara Aknin and her colleagues. She has tested the relationship between giving to others and personal well-being in samples of university undergraduates, children, and company employees; using vast international surveys; using lab and field experiments; and run in various countries including Canada, Uganda, India, South Africa, and a tiny village in Vanuatu (see Aknin et al., 2013; Aknin, Broesch, Hamlin, & Van de Vondervoort, 2015; Aknin, Hamlin, & Dunn, 2012; Dunn, Aknin, & Norton, 2008). The results from this group replicate across all of these various groups, offering strong evidence that the positive impact of giving has high external validity: It generalizes widely to many different populations.

The Internet is one relatively inexpensive way for researchers to reach samples beyond undergraduate students. Although online samples raise their own issues of generalization, they typically reach a broader population than undergraduate samples (Gosling, Vazire, Srivastava, & John, 2004). Online samples can be recruited using social media, or through companies like Amazon’s Mechanical Turk and Prolific Academic. These samples tend to have greater diversity with respect to socio-economic status, ethnicity, age, and work experience (Behrend, Sharek, Meade, & Wiebe, 2011; Casler, Bickel, & Hackett, 2013). Internet

samples also tend to produce rather similar responses as university undergraduates, thereby assuaging concerns that those completing studies online might be more prone to inattentive responding or other ways of producing invalid data (Bartneck, Duenser, Moltchanova, & Zawieska, 2015; Crump, McDonnell, & Gureckis, 2013). Continuing to seek evidence of external validity for our samples will always be an important goal.

Research with animals once suffered from the critique of relying too heavily on convenience samples, as well. At one time, animal research relied largely on the white rat, because rats were easy to obtain and study on a university campus (Beach, 1950). However, research with animals now relies on a great diversity of species to explore different research questions (Shettleworth, 2009). This animal research has often formed the basis of models for the biological underpinnings of cognition and perception, and these models have been successfully applied to humans: another demonstration of external validity. For example, research on reinforcement using rats and pigeons has been applied to humans to treat mental illness through behaviour modification, to understand personality traits, and to study decision-making.

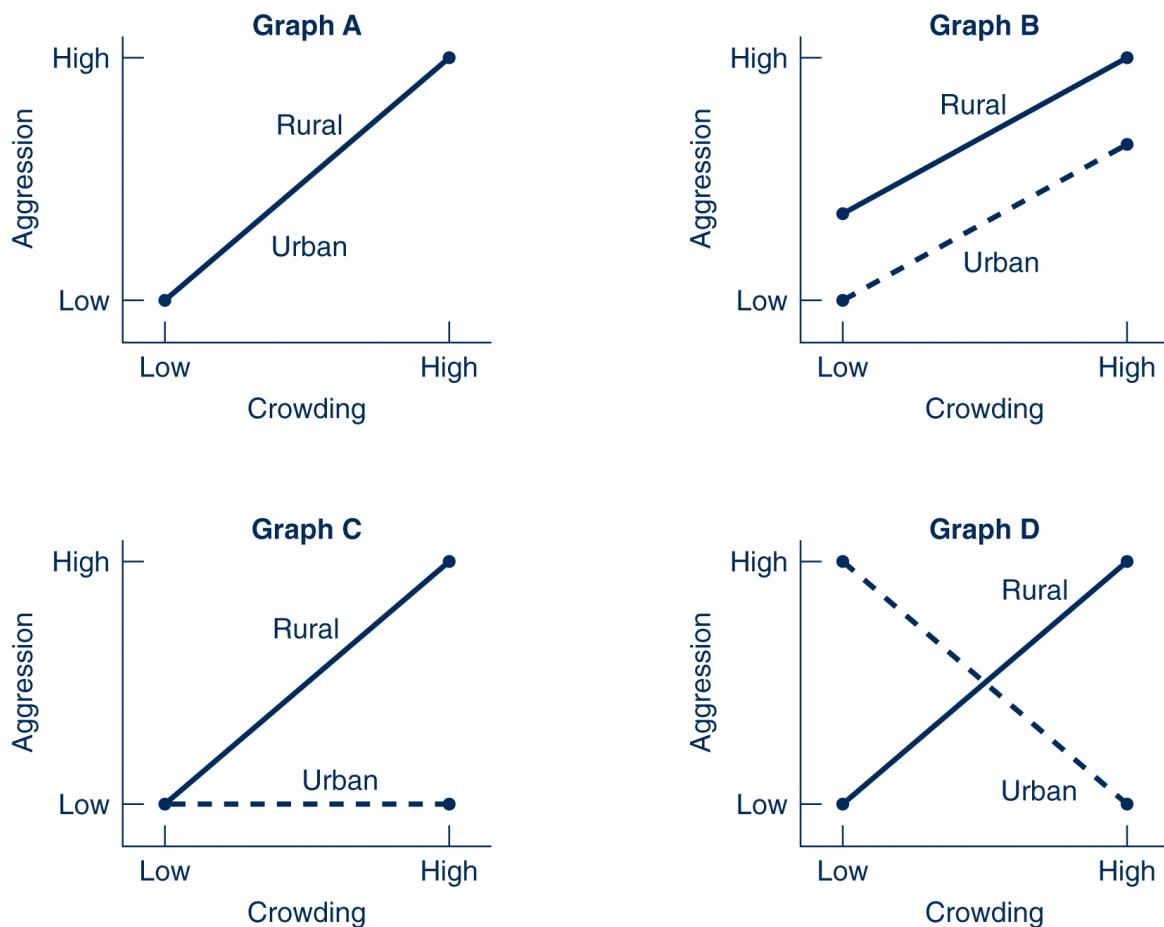
Examine the Influence of Group Membership Using Factorial Designs

Generalization to different populations can be studied directly using factorial designs ([Chapter 11](#)). Factorial designs allow us to study interactions, with an interaction occurring when an effect exists under one condition but not another, or when the nature of the effect is different in one condition than in another. This can be leveraged to study generalizability across groups directly. Imagine a study that used only rural participants, and you question its generalizability. What you are suggesting is that there might be an interaction between demographics and the independent variable. Let us consider a specific example. Imagine a study that examines the relationship between crowding and aggression, based only on rural participants: This study reports that crowding is associated with higher levels of aggression. You might then question whether the results are generalizable to those who were raised in an urban setting, within a crowded city.

[Figure 14.1](#) shows four potential outcomes of a hypothetical study on crowding and aggression that tested both urban and rural participants. In each graph, the relationship between crowding and aggression for rural participants has been maintained. In Graph A, there is no interaction—the behaviour of urban and rural participants is virtually identical (the solid and dotted lines overlap). Thus, the

results of the original study using rural participants can be generalized to urban dwellers. In Graph B, there is also no interaction; the effect of crowding is identical for urban and rural participants. However, in this graph, rural participants are more aggressive than urban ones (i.e., a main effect of being from a rural context). Although such a main effect of rural geography may be interesting, it does not prevent generalization because the overall positive relationship between crowding and aggression is present for both urban and rural participants.

Figure 14.1 Outcomes of a hypothetical experiment on crowding and aggression



Note: The presence of an interaction (Graphs C and D) indicates that the results for rural participants *cannot* be generalized to urban participants.



Graphs C and D show interactions, which demonstrate a problem for generalization. In both, the results with rural participants cannot be generalized to urban participants. In Graph C, there is no relationship between crowding and aggression for urban participants but there is for rural participants. In Graph D, the interaction tells us that a positive relationship between crowding and aggression exists for rural participants, but that the exact opposite relations, a negative relationship, exists for urban participants.

Researchers can address generalization issues that stem from the use of different populations by including individual differences as a variable in a factorial design. By including variables such as age or cultural background in the design of a study, it can be determined whether there are interaction effects like the ones illustrated in [Figure 14.1](#).Page 288



LO5 Rely on Multiple Studies to Draw Conclusions: Literature Reviews and Meta-analyses

In the past, researchers often drew conclusions about the generalizability of research findings by conducting literature reviews. In a [*literature review*](#), a researcher reads many studies that address a particular topic and then writes a paper that summarizes, organizes, and evaluates the literature, sometimes proposing advances to theory. The literature review offers a summary of existing results, indicating what findings are strongly or weakly supported in the literature, identifying inconsistent findings and areas in which research is lacking, and discussing future directions for research. The conclusions in a narrative literature review are based on the subjective impressions of the author.

A currently more prevalent approach to the review paper is to use statistical techniques to combine the results from many past studies, known as a [meta-analysis](#) (Rosenthal, 1991). In a meta-analysis, information from each past study is gathered and considered statistically, including effect-sizes, sample sizes, and other features of the studies that might influence the effect-size for the outcome (e.g., what population was studied, whether the study was published or remained unpublished). A meta-analysis compares the strength of a given finding across many different studies that tested the same or similar variables, in order to examine the best estimate of the effect-size for a given effect. A good meta-analysis can be extremely informative, and often serves to reconcile mixed findings in a research literature (i.e., some studies finding an effect, whereas others do not). The example of a meta-analysis on the so-called Mozart Effect is a good example of this (Pietschnig et al., 2010). One advantage of meta-analyses over literature reviews is that a single conclusion is evident, in the form of an estimation of the effect-size observed across many past studies. With a literature review, it is very difficult to integrate so many past results and arrive at a clear conclusion. In contrast, a meta-analysis using a quantitative approach, analyzing the data with statistics, to arrive at an estimate that integrates all this past information. However, because meta-analyses are quantitative in nature, they can only summarize and integrate the results from past quantitative studies; qualitative studies are excluded by necessity.

Meta-analyses are used to evaluate the relationship between variables, to test hypotheses, and to refine theories, all by incorporating the available quantitative research on a topic (Chan & Arvey, 2012). For example, Piers Steel (2007) from the University of Calgary used a meta-analysis to investigate the predictors and outcomes of procrastination. A total of 216 studies were found that investigated this topic, reporting a total of 691 different correlation coefficients. This meta-analysis found that people's tendency to procrastinate was related to the personality trait conscientiousness (and similar concepts like organization and low distractibility), but not to neuroticism (or similar concepts like perfectionism). In other words, in general, people tend to procrastinate because they are disorganized and impulsive, but less so because they are anxious or perfectionistic. Moreover, people tend to procrastinate on tasks they don't enjoy and when the reward for completing a task is not immediately delivered.

Meta-analytic results are typically displayed in tables, which contain information about the effect-size obtained in each of the past studies considered, along with a summary of the average effect-size across these studies. In [Table 14.1](#), we see some results from a meta-analysis on what predicts academic performance in

university, based on 241 separate datasets (Richardson, Abraham, & Bond, 2012). Performance self-efficacy was the strongest predictor of GPA, with an average r of .59, calculated across four samples encompassing over 1,300 participants. Notice the 95 percent confidence interval around this average correlation, indicating our degree of uncertainty around this estimated value. By pooling the results from dozens of past studies, often from various researchers working in different countries, meta-analysis can provide a more informative estimate of population values compared to relying on a single study.

Table 14.1 Results from a meta-analysis on predictors of undergraduate grade point average (GPA)

Correlate of GPA	Effect-Size r	95% Confidence Interval for r	Number of Samples	Total Sample Size
Socio-economic status	.11	[.08, .15]	21	75,000
High school GPA	.40	[.35, .45]	46	34,724
Trait conscientiousness	.19	[.17, .22]	69	27,875
Trait extraversion	-.04	[-.07, -.02]	58	23,730
Performance self-efficacy (belief in one's ability to succeed on familiar challenges)	.59	[.49, .67]	4	1,348
Academic self-efficacy (belief in one's ability to succeed on unfamiliar challenges)	.31	[.28, .34]	67	46,570
Academic intrinsic motivation	.17	[.12, .23]	22	7,414
Academic extrinsic motivation	.01	[-.06, .08]	10	2,339
Test anxiety	-.24	[-.29, -.20]	29	13,497
Goal commitment	.15	[.07, .22]	10	13,098
General stress	-.13	[-.19, -.06]	8	1,736
Depression	-.10	[-.17, .02]	17	6,335

Both narrative literature reviews and meta-analyses provide valuable information, and the two are often complementary. A meta-analysis allows statistical,

quantitative conclusions, whereas the narrative review uses a more qualitative approach to help identify trends in the literature and directions for future study. However, the quality of conclusions drawn from either method depends on the accuracy of the literature included (Chan & Arvey, 2012). The goal of anyone doing a meta-analysis or review is to consider all of the relevant research, whether it is published or not. This goal is complicated by publication bias (the tendency to publish only statistically significant results), which can lead to overestimating effect-sizes ([Chapter 13](#)). Therefore, researchers conducting these types of reviews often e-mail other researchers, or make posts to forums and listservs, requesting relevant unpublished data.

Even without conducting a meta-analysis ourselves, a background in meta-analysis is helpful when reviewing research findings. Simply knowing about meta-analysis can improve the way we interpret information for literature reviews, enabling us to discount flashy titles and focus on the actual effects, to make more accurate interpretations (Bushman & Wells, 2001).

The issues surrounding generalization for research results are complex and important. Replication, especially direct replication, and meta-analyses are now increasingly recognized as vital tools for generalizing beyond individual studies to achieve more accurate estimates of effects. Whenever possible, seek large and diverse samples for participation in research, to enable more accurate and generalizable results.



LO6 Generalizing Your Knowledge beyond This Book

This last chapter has emphasized the ability to generalize results from a study to different populations. As you finish reading this book, we encourage you to pause and consider ways you can “generalize” your knowledge of research methods beyond this course. This section is intended to help you identify some of ways of doing just that.

Recognize and Use Your New Knowledge

In [Chapter 1](#) we considered an overview of the research process. Examine [Figure 14.2](#), which connects these different steps of the research process to chapters of this book. Notice how we have emphasized study design across so many different chapters. As you likely noticed, there are many features to consider, and details to be decided, when designing methodologically rigorous research. To the extent you have engaged with this book (and listened to your course instructor), you have deepened your skills and understanding of proper research design. Not only can you use this knowledge to conduct high-quality research yourself, you can also use this knowledge to more effectively evaluate research reported in the media. This will also be helpful when you learn about research findings in future

courses. With this knowledge you can also seek out research findings to help you answer questions that concern you, such as “What are the best study strategies?” (Dunlosky et al., 2013), and “Do ‘dog people’ differ from ‘cat people’?” (Gosling, Sandy, & Potter, 2010). You might also be interested in becoming more involved in doing research, seeking further research training, and perhaps pursuing a career path that relies on these research skills ([Chapter 1](#)).

Figure 14.2 Overview of research process, with connections across chapters



Replicate or Extend Results | 14

Write Manuscript | 2 & A

Submit Manuscript to
Peer-Reviewed Journal | A



Stay Connected to Building a Better Psychological Science

One goal of this text was to address the current controversies, debates, and disciplinary reform initiatives related to research methods. These ongoing debates are particularly relevant to ethics ([Chapter 3](#)), statistics ([Chapters 12 and 13](#)), and generalization (this chapter). What constitutes the best practices in research methods continue to be shaped by active researchers in the field. The *Association for Psychological Science*, and particularly its journal *Perspectives on Psychological Science*, is a good place to keep current with these ongoing issues (www.psychologicalscience.org). It is an exciting time to be a psychological scientist!Page 291

In addition, you have now developed skills useful for accurately communicating research results to the public. Mahzarin Banaji, when she was president of the Association for Psychological Science, launched the *Wikipedia Initiative* “to improve the quality and quantity of the information about psychological science presented in Wikipedia” (Banaji, 2011; see also Banaji, 2010). By challenging all of us who are knowledgeable about psychological science—including students like you!—to update and clarify Wikipedia entries, more accurate information is reaching the broader public, to the benefit of all (Farzan & Kraut, 2013).

Use Research to Improve Lives

One reason why many students are drawn to psychology is the desire to help others. The topics we study and methods we use in our discipline make it well-positioned to “promote human welfare” (Miller, 1969) and to improve everyday life. Psychological research has indeed impacted many domains of life (Zimbardo, 2004). This impact can be seen in areas such as health (e.g., promoting healthy behaviours, reducing stress), law and criminal justice (e.g., understanding jury decision-making, improving eyewitness identification), education (e.g., improving academic performance), and work environments (e.g., motivating workers).

Psychologists have developed many websites to provide the public with information on parenting, education, mental health, and many other topics. Spread the word to your friends and families about ways to find research-based information to use in everyday life. For examples, share the websites of the American Psychological Association (www.apa.org), the Association for Psychological Science (www.psychologicalscience.org), the Canadian Psychological Association (www.cpa.ca), the Canadian Mental Health Association (www.cmha.ca), and the Greater Good Science Center (<http://greatergood.berkeley.edu/>). In addition, journals such as *Psychological Science in the Public Interest* present current reviews of topics that interest the public. Other journals present short reviews of emerging topics that might interest a junior researcher, such as *Current Directions in Psychological Science*.

Despite the potential challenges of generalizing research findings that have been highlighted in this chapter, evidence suggests that research findings can be used to improve the lives of many different groups. We hope that by engaging with this textbook, you have gained skills and knowledge that you can carry and apply beyond this course. Continue to use your skills to seek out, evaluate, and perhaps even create research evidence, so you can improve your life and the lives of others.



Illustrative Article: Generalizing Results

Driving around in a 4,000-pound automobile is a dangerous thing. Motor vehicle accidents are among the leading preventable causes of death in the United States every year. Distraction is one of the most common causes of automobile accidents, and talking to another person is a very common distraction.

In an effort to observe the impact of conversation on driving, Drews, Pasupathi, and Strayer (2008) conducted a study using a driving simulator that tracks errors committed by drivers. The researchers varied the type of conversation. In one condition, participants had a conversation with a passenger; in another condition, participants talked on a cellphone. There was also a no conversation, control condition. As you would expect, having any conversation resulted in more driving errors. However, the number of driving errors was highest in the cellphone condition. Page 292

For this exercise, acquire and read the article:

- Drews, F., Pasupathi, M., & Strayer, D. (2008). Passenger and cell phone conversations in simulated driving. *Journal of Experimental Psychology: Applied*, 14, 392–400. doi:10.1037/a0013119

After reading the article, consider the following:

1. Describe how well you think the sample of participants in this study generalizes to other groups of people.
2. In this study, participants were told to have a conversation about a time when “their lives were threatened.” Do you think that the results of this study would be different if the conversation were about something else? How so? Why?
3. Do you think that the findings from this study would generalize to other cultures? Do you think that samples of college and university

students in Mexico, Italy, and Germany would generate similar results? Why or why not?

4. How well do you think the driving simulator generalizes to real-world driving? What would you change to improve the generalizability of the simulator?
5. Evaluate the internal validity of this study. Explain your answer.
6. Evaluate the external validity of this study. Explain your answer.

Study Terms

Test yourself! Define and generate an example of each of these key terms.

- *conceptual replication* (p. 284).
- *direct replication* (p. 281).
- *external validity* (p. 277).
- *literature review* (p. 288).
- *meta-analysis* (p. 288).
- *replication* (p. 281).

Review Questions

Test yourself on this chapter's learning objectives. Can you answer each of these questions?

1. Why should a researcher be concerned about generalizing to other participant populations? What are some of the generalization problems that a researcher might confront?
2. What is the source of the problem of generalizing to other experimenters? How can this problem be solved?
3. Why might a pretest present a problem for generalization?
4. Distinguish between a direct replication and a conceptual replication. What is the value of each?Page 293
5. What is a meta-analysis? What is the purpose of a meta-analysis? How does it differ from a literature review?
6. List three ways you might continue to use your knowledge of research methods in the future.

Deepen Your Understanding

Develop your mastery of these concepts by considering these application questions. Compare your responses with those from other people in your study group.

1. Participate in a psychological research study on the Internet (e.g., <https://implicit.harvard.edu/implicit/>). What issues of generalization might arise when interpreting the results of such studies? Does the online aspect of the research make this research less generalizable than lab-based research? Or does the fact that people throughout the world can participate make it more generalizable? How could you design a study to answer this question empirically?
2. Use *PsycINFO* to find abstracts of articles that included culture or nationality as a variable (consult [Appendix D](#) for strategies). What conclusions do the authors of these studies draw about generalization?
3. Find a meta-analysis published in a journal. Two good sources are *Psychological Bulletin* and the *Personality and Social Psychology Review*. How were studies selected for the analysis? How was the concept of effect-size discussed in the meta-analysis? What conclusions were drawn from the meta-analysis?
4. Find the most recent Registered Replication Report in *Perspectives on Psychological Science*, *Advances in Methods and Practices in Psychological Science*, *Comprehensive Results in Social Psychology*, or another direct replication study. How many labs and participants were involved? How were issues of external validity and overall effect-size addressed? What conclusions were drawn about the original study results?

Appendix A

Writing Research Reports in APA Style

This appendix offers an introduction to reporting research results using the conventions of APA style. Most recommendations—including a complete, annotated example—relate to preparing written reports; we also include a few tips for presenting posters and talks. Information in this appendix can be helpful for writing or otherwise presenting your research results in your courses, at conferences, or for possible publication in a professional journal. When preparing reports for a class, a thesis, or submission to a journal, always check whether you are required to follow APA style directly, or other variations.

The format presented here for writing research reports is drawn from the *Publication Manual of the American Psychological Association* (Sixth Edition, 2010). *APA style* is the term used to denote this set of specific rules for organizing and presenting research results, as well as for citing and referencing past research. APA style is considered the standard in many journals in psychology, child development, family relations, and education, regardless of where you conduct your research. Other disciplines use other styles (e.g., MLA, Chicago). (If you are unsure whether a particular journal uses APA style, examine a recent issue of that journal.) To deepen your knowledge of APA style beyond the brief treatment presented here, purchase the entire *Publication Manual* through your university bookstore or directly from the American Psychological Association. Tutorials at www.apastyle.org may also be helpful.

Overall, the APA manual is guided by principles of “specificity and sensitivity.” First, papers should be written at a level of specificity and detail that will allow others to *replicate* the research. Second, papers should

be free of inappropriate language that might be interpreted as biased and insensitive. The manual also includes manuscript preparation guidelines, including references. Throughout this appendix, examples that are intended to appear as you would type them in your papers appear in a unique font to make spacing and other rules clear; this convention is also used in the APA manual. When typing your own paper, you would not use this type of font (see [below](#) for details).

Why learn APA style? The purpose of APA style is to facilitate communication between writers and readers. Because APA style provides a uniform structure for written reports and other presentations, you will be better equipped to understand and evaluate them. By using APA style properly in your own work, you signal that you are a participant in this scholarly community while making it easier for other members of this community to understand your work. As you will notice throughout this appendix, attention to detail is crucial to mastering these rules. Mastering APA style provides an opportunity to develop your detail skills, which will prove helpful when you enter the job market.

Many additional resources might be helpful as you prepare to write your first research reports. In a brief book, former *Psychological Science* editor Robert Kail (2015) offers specific advice on how to effectively construct sentences, paragraphs, and APA style manuscripts. It serves as a tutorial by incorporating many examples, prompts to practise, and answers. For additional guidelines about writing compelling APA style research reports, see an article by Kazdin (1995) and a chapter by Bem (2003). Other brief books offer advice spanning many aspects of report writing, from conducting literature reviews through to preparing manuscripts for publication (see Rosnow & Rosnow, 2012; Sternberg & Sternberg, 2010).Page 295

The rest of this appendix is divided into sections: general writing style and word choice, writing each section of the APA style research report, formatting a manuscript, citing and referencing sources in APA style, and tips for conference presentations. Finally, an example paper is provided in manuscript format ready for submission to a journal.

Writing Style and Word Choice

One way to think about writing style is as a way to signal membership in a community. Think about how you might communicate the same idea—say, about how prepared you are to write tomorrow’s exam—differently depending on whether you are talking to your friend, a parent, or the course instructor. You might choose different words, emphasize different points, and omit different information. Similarly, writing about research effectively requires adapting your writing to fit in with the style of the academic community (e.g., psychology). Writing style that is appropriate for the psychology community would seem out of place in the history community, and vice versa. Learning to write effectively in the style of a discipline involves identifying and using subtle features that indicate membership in that community (Madigan, Johnson, & Linton, 1995).

Identifying and learning the subtle stylistic features of a discipline can be difficult, precisely because they are so subtle. Whenever you read published journal articles, try to notice how the authors choose to incorporate past research, summarize research results, and discuss the implications of their work. We provide some general suggestions for improving writing, and as much as possible tailor recommendations to the style of psychology. We recommend consulting Madigan et al. (1995) and Kail (2015) for further insight into the subtleties of this style.

Clarity and Intended Audience

Present ideas precisely and clearly, with your intended audience in mind. Words are a vehicle for communicating ideas. It is important to use straightforward sentences that avoid flowery language. Some students seem to believe that liberal use of a thesaurus and extremely long complex sentences makes their work sound smart. Instead, the sacrifice in clarity signals that the author is not a member of this scholarly community. As people become more effective writers in psychology, they are able to

maintain clarity while injecting creativity into their reports. But clarity is crucial for writing effectively in this community.

The overall piece should be organized in such a way that the ideas flow coherently and logically. APA style provides an overall organizational structure for research reports that can also be used in other forms of professional communication (e.g., posters). Yet each section within the APA style written report needs to be organized by the author (particularly the introduction and discussion sections, see [below](#)). Creating an outline is one way to organize sections (as well as other works that do not use APA style). Many writers plan a paper by putting their thoughts and ideas into outline form. The outline then serves as a writing guide. This method forces writers to develop a logical structure before writing the paper.

Paragraphs should be well organized. It is a good idea for a paragraph to contain a topic sentence. Other sentences within a paragraph should develop the idea in the topic sentence by elaborating, expanding, explaining, or supporting the same idea. Kail (2015) recommends developing the outline for subsections of the report from topic sentences, then using those topic sentences to guide your writing of the rest of the paper. Avoid one-sentence paragraphs. If you find such paragraphs in your paper, expand the paragraph, include the idea in another paragraph, or delete the concept.[Page 296](#)

The amount of jargon and level of specificity you use will depend on your audience. When preparing any kind of writing in a class, ask your instructor whether the audience is other scholars knowledgeable about the topic and methods, other scholars unfamiliar with the topic and methods, the general public, or some other group. Use only the jargon that is appropriate for the audience you are intending to reach. For example, it would be appropriate to use the term *external validity* when writing a report for publication in a journal, but might be less appropriate when writing a blog post or a Wikipedia entry. In reports for publication, assume the reader is generally familiar with statistics and hypothesis testing. Statistical outcomes can usually be presented without defining terms such as the *mean*, *standard deviation*, or *significance*. Regardless of your audience, avoid creating

unnecessary abbreviations your readers have to learn (e.g., *LM condition* to refer to the *laptop multitasking condition*).

Expect to write multiple drafts of your paper. After completing the first draft of your paper, it is a good idea to let it sit for a day or so before you reread it. Carefully proofread the paper, paying attention to grammar and spelling. Some grammatical considerations are described here; you can also use your word processor to check your spelling and grammar. We advise seeking feedback from others who will read your report carefully and suggest improvements. Ideally, these readers are knowledgeable with the type of writing you are preparing (e.g., APA style report) but not necessarily the topic. Thoughtfully consider what feedback will improve your final product.

Paraphrase and Cite Past Research

There are two important reasons to cite past research in your report. First, it connects your research to the rest of what is known about a topic.

Acknowledging what other researchers have already found enables you to specify the contribution your research makes to knowledge (Giltrow, Gooding, Burgoyne, & Sawatsky, 2014). Second, acknowledging the work of others signals what ideas and results are theirs, and what ideas and results are yours. In other words, you avoid plagiarism (Chapter 3).

In psychology and other sciences, it is common to include others' work by paraphrasing it and adding a citation at the end of the sentence.

Paraphrasing means re-stating the original author's idea or result in different words. Learning to paraphrase can be challenging, particularly for students who are still building confidence in writing. Examples of effective paraphrasing are available online (e.g., see

<http://www.uc.utoronto.ca/paraphrase> and https://owl.purdue.edu/owl/-research_and_citation/using_research/quoting_paraphrasing_and-summarizing/paraphrasing_sample_essay.html for practice exercises).

Practise paraphrasing others' work as you use it to identify the contributions you are making. Indicating how your paper builds on previous research strengthens your paper by showing the reader that you are building on the existing body of scientific knowledge (Harris, 2002).

Refer to the section on citing and referencing sources [later](#) in this appendix for details about how to cite using APA style.

Paraphrasing is a stylistic feature that psychology shares with other sciences, but is less common in the humanities (Madigan et al., 1995). If you have taken many courses in the humanities, you may be used to directly quoting your sources. The tendency to paraphrase rather than quote in the sciences reflects an assumption that an idea can be preserved regardless of the precise words used. To signal that you are participating in the scientific community, avoid using direct quotations and instead use paraphrasing with citations. Consult your instructor or the writing centre at your university for more advice on effective and honest use of sources to write in psychology.^{Page 297}

Active versus Passive Voice

It is common for authors writing in psychology to use the passive voice (Madigan et al., 1995). Some argue that many writers rely too much on the passive voice in their reports, and risk lost clarity (Kail, 2015). Consider the following sentences:

- It was found by Yee and Johnson (1996) that adolescents prefer . . .
- Participants were administered the test after a 10-minute rest period.
- Participants were read the instructions by the experimenter.

Now try writing those sentences in a more active voice. For example:

- Adolescents prefer . . . (Yee & Johnson, 1996).
- Participants took the test after a 10-minute rest period.
- The experimenter read the instructions to participants.

Prose that seems stilted using the passive voice is much more direct when phrased in the active voice. Deliberately choose active or passive voice.

Sometimes authors refer to themselves in the third person. Thus, they might say “The experimenter distributed the questionnaires” instead of “I distributed the questionnaires.” It is unclear whether “the experimenter” is in fact the author or someone else. When authors refer to themselves in the paper, APA style recommends using first person pronouns (e.g., “I” or “we”).

Avoiding Biased Language

Recall that APA style is guided by the principles of specificity and sensitivity. The principle of specificity leads to the recommendation to use the term *participants* to refer to humans and *subjects* to refer to animals used in your study. Use other terms if they more accurately describe the people in the study (e.g., children, patients, clients, or *respondents* in survey research).

Be sensitive to the possibility that your writing might convey a bias, however unintentional, regarding gender, sexual orientation, and ethnic or racial group. As a general principle, be as specific as possible when referring to groups of people. For example, referring to the participants in your study as “Korean Canadians and Vietnamese Canadians” is more specific and accurate than describing them as “Asians.” Also, be sensitive to the use of labels that might be offensive to members of certain groups. In practice, this means using the terms people prefer. Instead of writing “We tested groups of schizophrenics and normals,” write “We tested people with and without schizophrenia.”

The APA manual offers numerous examples of ways to be sensitive to gender, racial and ethnic identity, age, sexual orientation, and disabilities. The term *gender* refers to males and females as social groups. Thus, gender is the proper term to use in a phrase such as “gender difference in average salary.” The term *sex* refers to biological aspects of men and women; for example, “sex fantasies” or “sex differences in the size of certain brain structures.” The use of gender pronouns can be problematic. Do not use *he*, *his*, *man*, *man's*, and so on when both males and females are meant. Sentences can usually be rephrased or specific pronouns deleted to avoid linguistic biases. For example, “The worker is paid according to his

productivity” can be changed to “The worker is paid according to productivity” or “Workers are paid according to their productivity.” In the first case, *his* was simply deleted; in the second case, the subject of the sentence was changed to plural. Avoid substituting pronouns with *s/he*.Page 298

There are certain rules to follow when referring to racial and ethnic groups. The names of these groups are capitalized and never hyphenated; for example, Black, White, African American, Latino, South Asian, Chinese Canadian. The manual also reminds us that the terms that members of racial and ethnic groups use to describe themselves may change over time, and there may be a lack of consensus about a preferred term. You are urged to use the term most preferred by your participants. If you have any questions about appropriate language, consult the APA style manual and your instructor or colleagues whose opinions you respect.

Writing Each Section of the APA Style Research Report

A research report is organized into six major parts: Abstract, Introduction, Method, Results, Discussion, and References. The report may also include tables and figures used in presenting the results. We will consider the parts of the paper in the order prescribed by APA style, but you do not have to write them in this order. Consider starting with the Method section because it is relatively easy to describe what participants did. Refer to the [sample paper](#) at the end of this appendix as you read the material that follows. As you read the published literature, you will see variations on this basic structure to account for multiple studies in a single paper, meta-analyses, and literature reviews.

Title Page

The first page of the paper is the title page. It is a separate page and is numbered page 1. Note that in the [sample paper](#), the title page includes the title as well as other important information. The first line of the title page is the *running head* and a page number (1). The running head has a very specific meaning and purpose: It is an abbreviated title and should be no more than 50 characters (letters, numbers, spaces) long. The running head line from the example article appears as follows:

Running head: GETTING A BIGGER SLICE

Note that all letters in the running head are capitalized, but only the R in Running head is capitalized (not the h). If the paper is published in a journal, the running head is printed as a heading at the top of pages (along with the page number) to help readers identify the article.

The running head and page number should be formatted so that the page number is flush to the right margin of the paper and the running head is

flush to the left margin of the paper. Do not try to manually type the running head and number at the top of every page of your paper; check your word processing program for how to create a page header. Use the page header feature to create a header that prints approximately halfway between the text of the paper and the top of each page, usually 0.5 inch (1.25 cm) from the top. The same page header (including the running head and page number) appears on every page of your paper.

The remainder of the title page consists of the title, author, and institutional affiliation. All are centred on the page. The title should be fairly short (usually no more than 10 to 12 words) and should inform the reader of the nature of your research. A good way to do this is to include the names of your variables in the title. For example, the following titles are both short and informative:

- Anxiety Impairs Mathematical Problem Solving
- Laptop Multitasking Hinders Classroom Learning for Both Users and Nearby PeersPage 299

Sometimes a colon in the title will help to convey the nature of your research or even add a bit of “flair” to your title, as in the following:

- Cognitive Responses in Persuasion: Affective and Evaluative Determinants
- The Pen Is Mightier Than the Keyboard: Advantages of Longhand over Laptop Note-Taking

Another method of titling a paper is to pose the question that the research addresses, as in these examples:

- Do Rewards in the Classroom Undermine Intrinsic Motivation?
- Does Occupational Stereotyping Still Exist?

Search engines (e.g., *PsycINFO*) are most likely to find your article if the title includes words and phrases that people are most likely to use when

conducting a search on your topic. This consideration also applies to the abstract.

Abstract

The abstract is a brief summary of the research report and typically runs 100 to 120 words in length. The purpose of the abstract is to introduce the article, allowing readers to decide whether the article appears relevant to their own interests. The abstract should provide enough information so that the reader can decide whether to read the entire report, and it should make the report easier to comprehend when it is read.

Although the abstract appears at the beginning of your report, it is easiest to write the abstract last. Read a few abstracts to get some good ideas for how to condense a full-length research report down to eight or ten information-packed sentences. For practice, write an abstract for a published article, and then compare your abstract to the one written by the original authors.

Abstracts generally include a sentence or two about each of the four main sections in the body of the article. First, from the Introduction section, state the problem under study and the primary hypotheses. Second, from the Method section, include a brief summary of the procedure (e.g., self-report questionnaires, direct observation, repeated measurements), and possibly some information on participants' characteristics (e.g., number, age, sex, and any special characteristics). Third, from the Results section, describe the pattern of findings for major variables. This is typically done by reporting the direction of differences, omitting numerical values. Finally, the abstract will include implications of the study taken from the Discussion section. Informative comments about the findings are preferred to general statements such as "the implications of the study are addressed" (Kail, 2015; Kazdin, 1995).

The abstract is typed on a separate page numbered page 2. The word "Abstract" is centred at the top of the page. The abstract is always typed as a single paragraph with no paragraph indentation.

Introduction

The Introduction section begins on a new page (page 3), with the title of your report typed and centred at the top of the page. Note that the author's name does not appear on this page, which allows a reviewer to read the paper without knowing the name of the author. After reading the introduction, the reader should know why the research is important and how you decided to go about doing it. In general, the introduction progresses from broad theories and previous relevant research, to hypotheses and specifics of the current research.

The introduction has three components, although formal subsections are rarely used. The components are (1) the problem under study, (2) the literature review, and (3) the rationale and hypotheses of the study. The introduction should begin with an opening statement of the problem under study. In two or three sentences, give the reader an appreciation of the broad context and significance of the topic (Bem, 1981; Kazdin, 1995). Specifically stating what problem is being investigated helps readers, even those who are unfamiliar with the topic, to understand and appreciate why the topic was studied in the first place.

Following the opening statement, the introduction describes past research and theory most relevant to your hypothesis. This is called the *literature review*. An exhaustive review of past theory and research is not necessary. (If there are major literature reviews of the topic, you would refer the reader to the reviews.) Rather, you want to describe only the research and theoretical issues that are clearly related to your study. State explicitly what is already known about your topic, and identify what is not known yet. By specifying such a gap in the existing knowledge, you are preparing readers to understand the contribution your study will make by filling that gap. Giltrow and colleagues (2014) offer a thorough discussion of the purpose of an introduction.

The final part of the introduction tells the reader the rationale of the current study. Here you state what variables you are studying and what results you expect. Your hypothesis and the current research design should follow

logically from your previous discussion of prior research and what knowledge is missing.

Method

The Method section begins immediately after you have completed the introduction (on the same page, if space permits). This section provides the reader with detailed information about how your study was conducted. Ideally, there should be enough information in the Method section to allow a reader to directly replicate your study.

The Method section is typically divided into subsections. Both the order and the number of subsections vary in published articles. Decisions about which subsections to include are guided by the complexity of the investigation. The [sample paper](#) in this appendix uses three subsections: Participants, Materials, and Procedure. Some of the most commonly used subsections are discussed next.

Overview

If the experimental design and procedures used in the research are complex, a brief overview of the method can help the reader understand the information that follows.

Participants

A subsection describing the participants is always necessary. Include the number of participants, as well as relevant characteristics such as age, sex, and ethnicity. Include any special characteristics that are relevant to your research question. For example, you may have limited your sample to first-born children, adolescent children of alcoholics, student teachers, or parents of children being treated for ADHD. State explicitly how participants were recruited and what incentives for participation were used, if any.

Apparatus

An Apparatus subsection may be necessary if special equipment is used in the experiment (e.g., an eye-tracking device). Specify the brand name and model number of special equipment. If the device is rare or has never been used for research before, consider describing it in detail. Include this information if it is needed to replicate the study.

Procedure

The Procedure subsection tells the reader exactly how the study was conducted. Include any detail that might be important in a direct replication of the study. One way to report this information is to describe, step by step, what occurred in the study from the perspective of the participant. Maintain the temporal sequence of events so that the reader is able to visualize the sequence of events the participants experienced.

The Procedure subsection tells the reader what instructions were given to the participants, how the independent variables were manipulated, and how the dependent variables were measured. The methods used to control extraneous variables also should be described. These include random assignment procedures, counter-balancing, and any special means that were used to keep a variable constant across all conditions. Describe how participants were debriefed, particularly if deception was used. If your study used a non-experimental method, provide details on exactly how you conducted the study and all measures you used. Ask a colleague to read your procedure to ensure it is appropriately detailed and clear.

Other Subsections

Include other subsections if they are needed to clearly present the method. For example, a subsection on testing materials (e.g., questionnaires) might be necessary instead of an Apparatus subsection. Other sections are customized by the authors to suit their study. If you glance through a recent issue of a journal, you will find that some studies have only two subsections and others have many more subsections. This reflects the varying complexity of the studies and the particular writing styles of the researchers.

Results

The Results section is a straightforward description of your results, supported by appropriate statistical analyses. Although it is tempting to explain your findings in the Results section, save that discussion for the next section of the paper. Focus on presenting the results as clearly and efficiently as possible.

The content of your Results section will vary according to the type and number analyses you conducted. If you stated more than one hypothesis in the Introduction section of the paper, consider presenting your results in the same order. If you conducted a manipulation check, consider presenting it before you describe the major results. Some authors include in the Results section a description of scoring or coding procedures performed on the data to prepare them for analysis. Other authors include such data transformations in a subsection of the Method section.

Summarize each finding in words, and use a statistical phrase at the end of the sentence to indicate the type of statistical test used to draw that conclusion. These statistical phrases indicate the type of test used, the degrees of freedom, the exact *p*-value, and the effect-size (Chapter 13). Reserve the term “significant” to refer to findings that have a *p*-value less than your alpha. Your readers will assume that you used an alpha (probability) level of .05 for decisions about statistical significance. If you did not, add a simple sentence such as “An alpha level of .01 was used for statistical analyses.”

Report statistical values (e.g., mean, standard deviation, *t*) to two decimal places. Also, round probabilities to two decimals (e.g., *p* = .03); any value less than .01 should be reported as *p* < .01. This guideline has not been universally adopted, so you will read many articles in which statistics and/or probabilities are reported using three decimal places. Your instructor may require the use of three decimal places. Page 302

The results should be stated in simple sentences. For example, consider the difference in life satisfaction among Australian and Mexican respondents we considered in Chapter 12. We could express that difference as follows:

Contrary to predictions, Australian respondents reported lower life satisfaction ($M = 7.20$, $SD = 2.05$) than Mexican respondents ($M = 8.51$, $SD = 1.93$), $t(3463) = 19.19$, $p < .01$, Cohen's $d = .66$.

These brief sentences inform the reader of the general patterns of the results, the obtained means, statistical significance, and effect-size. Carefully note the precision of the statistical phrase. Each space (e.g., before and after =), comma, and the order that values are presented is specific to APA style.

If the results are relatively straightforward, they can be presented entirely in sentence form. If the study involved a complex design, use tables and figures to present your results clearly.

Tables and Figures

Tables are generally used to present large arrays of data. For example, a table might be useful in a design with several dependent measures; the means of the different groups for all dependent measures would be presented in the table. Tables are also convenient when a factorial design has been used. For example, in a $2 \times 2 \times 3$ factorial design, a table could be used to present all 12 cell means.

Figures are used when a visual display of the results would help the reader understand the outcome of the study, such as a significant interaction or a trend over time (Chapter 12). Bar graphs are used when describing the responses of two or more groups (e.g., experimental and control conditions, or Australian versus Mexican respondents). Line graphs are useful when both variables have quantitative properties, e.g., the average response time of two groups on days 1, 2, 3, 4, and 5 of an experiment. Consult Nicol and Pexman (2010) for detailed information on creating figures and other visual displays of data.

When strictly adhering to APA style, each table and figure appears on a separate page at the end of the manuscript, rather than presented in the main body. Check with your instructor for variations in course assignments and theses. In the Results section text, refer to the table or figure number

and briefly describe its main content. For example, make a statement such as “As shown in Figure 1, the laptop group . . .” or “Table 1 presents the demographic characteristics of the survey respondents.” Describe the important features of the table or figure rather than using a generic comment such as “See Figure 3.”

Avoid repeating the same data in more than one place. An informative table or figure supplements, not duplicates, the text. Using tables and figures does not diminish your responsibility to clearly state the nature of the results in the text of your report.

Discussion of the Results

It is usually *not* appropriate to discuss the implications of the results in the Results section. However, the Results and Discussion sections may be combined if the discussion is brief and greater clarity is achieved by the combination. This combination happens most often in papers with multiple studies. A General Discussion is then added at the end of all studies that aligns with the broader Discussion section described below.³⁰³

Discussion

The Discussion section is the proper place to discuss implications of the results. Just like the introduction, it is important for the Discussion section to be logically organized (see Kail, 2015, for further tips). One way to organize the discussion is to begin by summarizing the original purpose and expectations of the study, and then to state whether the results were consistent with your expectations. If the results do support your original ideas, you should discuss how your findings contribute to knowledge of the problem you investigated. You will want to consider the relationship between your results and past research and theory. If you did not obtain the expected results, discuss possible explanations. The explanations would be quite different depending on whether you obtained results that were the opposite of what you expected or the results were not significant.

It is often a good idea to include your own criticisms of the study. No study is ever perfect; all have limitations. It is appropriate to address any major

limitations in your study in the Discussion section. Try to anticipate what a reader might find wrong with your methodology. For example, if you used a non-experimental research design, you might acknowledge problems of cause and effect, and identify any specific possible extraneous variables you think might be operating. Sometimes there may be major or minor flaws that could be corrected in a subsequent study (if you had the time, money, and so on). You can describe such flaws and suggest corrections. You might argue whether the results would or would not generalize to other samples.

The results will probably have implications for future research. If so, discuss the direction that research might take. It is also possible that the results have practical implications—for example, for child-rearing or improving learning in the classroom. Discussion of these larger issues is usually placed at the end of the Discussion section. Finally, consider including a brief concluding paragraph that provides “closure” to the entire paper.

References

The list of references begins on a new page. The references must contain complete citations for all sources mentioned in your report. Do not omit any sources you cited from the list of references; also, do not include any sources that are not cited in your report. The exact procedures for citing sources within the body of your report and in your list of references follow the APA *Publication Manual* and are described [later](#) in this appendix. You can also follow examples in recent publications that use APA style.

Appendix

The APA *Publication Manual* notes that an appendix might be appropriate when necessary material would be distracting in the main body of the report. Examples of appendixes include the entire questionnaire or survey instrument, a complex mathematical proof, a long list of words used as stimulus items, or other materials employed in the study. If an appendix is provided in the manuscript itself, it begins on a new page with the word

“Appendix” centred at the top. Sometimes journals provide appendixes as online supplements to articles; other times, individual authors upload these materials to their websites. Check with your instructor concerning the appropriateness of an appendix for your paper.

Author Note

The author note begins with a paragraph that gives the department affiliations of the authors. Another paragraph may give details about the background of the study (e.g., that it is based on the first author’s master’s thesis) and acknowledgments (e.g., grant support, colleagues who assisted with the study, and so on). A final paragraph begins with “Correspondence concerning this article should be addressed to . . .” followed by the mailing and e-mail addresses of the author people should contact if they wish to follow up directly. The author note usually begins on a new page. However, sometimes a journal editor will ask you to place the author note information at the bottom of the title page. This is done when the paper will have a *masked review*: The person reviewing the paper has no information about the author of the paper. In this case, the title page will be separated from the rest of the paper prior to review. An author note may be unnecessary for class research reports; consult your instructor.

Footnotes and Endnotes

Footnotes are rarely used in psychology. In unpublished manuscripts, all footnotes in the paper are treated as endnotes, and are typed on one page at the end of the paper. Avoid using footnotes unless they are absolutely necessary. They can be distracting to readers, and important information can and should be integrated into the body of the paper.

Tables

Each table should be on a separate page. As noted previously, APA style requires placement of the table at the end of the paper, but for a class you may be asked to place your tables within the body of the paper. When preparing your table, allow enough space so that the table does not appear

cramped on a small portion of the page. Define areas of the table using horizontal lines (do not use vertical lines in APA style tables). Ensure that the title accurately and clearly describes the content of the table. You may wish to use an explanatory note at the bottom of the table to show significance levels or the range of possible values on a variable. There are common formats for many types of tables, including tables of means, correlation coefficients, multiple regression analyses, and so on (consult the *Publication Manual*; Nicol & Pexman, 2010; or a recent journal published by the APA). For example, below is a table of correlations. Note that the title of the table is typed in italics, and that the areas of the table are separated by horizontal lines.

Figures

According to APA style, figures are placed after the tables in papers. However, this rule may not be necessary for student reports or theses. You may be asked to place each figure on a separate page at the appropriate point in the body of the text. As on every page, the running head and page number appear at the top of each figure page. Page 305

Most spreadsheet, word processing, and statistical analysis programs have graphing features (e.g., Word, Excel, OpenOffice Calc, SPSS). Independent and predictor variables are placed on the horizontal axis; dependent and criterion variables are placed on the vertical axis. Both the horizontal and vertical axes must be labelled. When you print the graph on a separate piece of paper, a rule of thumb is that the horizontal axis should be about 5 inches (12.5 cm) wide, and the vertical axis should be about 3.5 inches (8.75 cm) long. If you are inserting a graph into the text of your report (not using APA style), your graphs may be smaller than this.

Remember that the purpose of a figure is to depict results clearly. If the graph is cluttered with information, it will confuse the reader and will not serve its purpose. Plan your graphs carefully to make sure that you are accurately and clearly informing the reader.

Summary: Order of Pages

To summarize, the organization of your paper is as follows:

1. Title page (page 1)
2. Abstract (page 2)
3. Pages of text (start on page 3)
 1. Title at top of first page begins the Introduction
 2. Method
 3. Results
 4. Discussion
4. References (start on new page)
5. Appendix (start on new page if included)
6. Author Note (start on new page)
7. Footnotes (start on new page if included)
8. Tables, with table captions (each table on a separate page)
9. Figures, with figure captions (each figure on a separate page)

You should now have a general idea of how to structure and write your report. The remainder of this appendix focuses on some of the technical rules that may be useful as you prepare your own research report.

Formatting a Manuscript

In APA style, the paper should be entirely double-spaced. The margins surrounding text should be 1 inch (2.5 cm) on all four sides of the page. Page headers, the information that appears at the top of each page including the page number, are set approximately 0.5 inch (1.25 cm) from the top of the page. All pages are numbered except for figure pages at the end of the paper. Paragraphs are indented 5 to 7 spaces or 0.5 inch (1.25 cm; use the tab keyboard function, not multiple spaces). Avoid using contractions (e.g., use *cannot* instead of *can't*). Rather than fully justifying text and breaking words with hyphens at the end of lines, justify your text to the left margin (creating a jagged edge of text along the right side of the page).

According to APA style, place only one space between sentences. Leaving two spaces is a convention left over from the days of manual typewriters; one space is more attractive and readable when using word processors with modern printer fonts. (If you automatically double-space after a period without thinking, use your word processor's "replace" feature to replace instances of two spaces with one space.)Page 306

Take advantage of the features of your word processing application. Learn to use headers to place running heads and page numbers automatically at the top of each page. Use other features to insert tables, centre text, check spelling and grammar, insert tabs (rather than spaces), and so on.

The font should be 12-point size throughout the paper. Use a serif font for all text and tables. The serif font should usually be Times New Roman font style. Figures, however, should be prepared with a sans serif font, either Arial or Calibri font style. Serif fonts have short lines at the ends of the strokes that form the letters; sans serif literally means "without serif" and so does not have serif lines. Here are examples:

This is serif text.

This is sans serif text.

Use the italics feature of your word processor sparingly. Italics are used for (a) titles and volume numbers of periodicals in the References section; (b) titles of books in the References section; (c) most statistical terms; (d) anchors of a scale, such as 1 (*strongly disagree*) to 5 (*strongly agree*); and (e) when you need to emphasize a particular word or phrase when first mentioned in the paper. Pay attention to the use of italics in the examples used throughout this appendix. Use boldface type to denote some headings, as noted below.

Using Headings

Papers written in APA style use one to five levels of headings. Most commonly, you will use level 1 and level 2 headings, and you may need to use level 3 headings as well. These five levels are as follows:

(Level 1) **Centred Heading**

(Level 2) **Margin Heading** The text begins indented on a new line.

(Level 3) **First paragraph heading.** The heading is bold, and indented as a new paragraph. The text begins on the same line.

(Level 4) **Second paragraph heading.** The heading is bold and italicized. It is indented as a new paragraph and the text begins on the same line.

(Level 5) **Third paragraph heading.** The heading is italicized (only). It is indented as a new paragraph and the text begins on the same line.

Level 1, or centred, headings are used to head major sections of the report: Abstract, Title (on page 3), Method, Results, Discussion, References. Level 1 headings are typed with uppercase and lowercase letters (i.e., the first letter of each major word is capitalized).

Level 2, or margin, headings are used to divide major sections into subsections. Level 2 headings are typed flush to the left margin, with

uppercase and lowercase letters (i.e., the first letter of each major word is capitalized). For example, the Method section is divided into at least two subsections: Participants and Procedure. The correct format is as follows:

Method

Participants

The description of the participants begins on a new line.

Procedure

The procedure is described in detail.

Page 307Level 3 to 5, or paragraph, headings are used to organize material within a subsection. For example, the Procedure subsection might be broken down into separate categories for describing instructions to participants, the independent variable manipulation, measurement of the dependent variable, and debriefing. Each of these could be introduced using level 3 paragraph headings. (For example, the Materials subsection in the [sample paper](#) at the end of this appendix uses three level 3 headings.) You may break these categories down further using heading levels 4 or 5, but these levels are rare.

Level 3 to 5 paragraph headings begin on a new line, indented 0.5 inch (1.25 cm). The first word begins with a capital letter; the remaining words are all typed in lowercase letters, except for proper nouns and the first word to follow a colon, which are capitalized. The heading ends with a period. All information that appears between a paragraph heading and the next heading (of any level) must be related to the paragraph heading. Both level 3 and level 4 headings are boldface; both level 4 and level 5 headings are italicized.

Abbreviations

Abbreviations are used sparingly in APA style papers. They can be distracting because the reader must constantly translate the abbreviation into its full meaning. However, APA style does allow for the use of abbreviations that are accepted as words in the dictionary (specifically, *Webster's Collegiate Dictionary*). These include IQ, REM, ESP, and AIDS.

Scientific abbreviations for various measurement units are also acceptable (e.g., cm for centimetre, ms for millisecond).

Certain well-known terms may be abbreviated when it would make reading easier, but the full meaning should be given when first used in the paper. Examples of commonly used abbreviations are the following:

MMPI	Minnesota Multiphasic Personality Inventory
STM	short-term memory
CS	conditioned stimulus
RT	reaction time
CVC	consonant-vowel-consonant
ANOVA	analysis of variance

Statistical terms are sometimes used in their abbreviated or symbol form. These are always italicized in a manuscript, as in the following examples:

<i>M</i>	mean
<i>SD</i>	standard deviation
<i>Mdn</i>	median
<i>df</i>	degrees of freedom
<i>n</i>	number of individuals in a group or experimental condition
<i>N</i>	total number of participants or respondents
<i>p</i>	probability (significance) level
<i>SS</i>	sum of squares
<i>MS</i>	mean square
<i>F</i>	value of <i>F</i> in analysis of variance
<i>r</i>	Pearson correlation coefficient
<i>R</i>	multiple correlation coefficient

Finally, certain abbreviations of Latin and Middle English terms are regularly used in papers. Some of these abbreviations and their meanings are given below:

- cf. compare (from Latin *confer*)
- e.g. for example (from Latin *exempli gratia*)
- et al. and others (from Latin *et alia*)
- etc. and so forth (from Latin *et cetera*)
- i.e. that is (from Latin *id est*)
- viz. namely
- vs. versus

Reporting Numbers and Statistics

Virtually all research papers report numbers: number of participants, number of groups, the values of statistics such as t , F or r . Should you use numbers (e.g., “43”), or should you use words (e.g., “forty-three”)? The general rule is to use words when expressing the numbers zero through nine but to use numbers for 10 and above. There are some important qualifications, however.

If you start a sentence with a number, you should use words even if the number is 10 or larger (e.g., “Eighty-five student teachers participated in the study.”). Starting a sentence with a number is often awkward, especially with large numbers. Therefore, you should try to revise the sentence to avoid the problem (e.g., “The participants were 85 students enrolled in teaching credential classes.”).

When numbers both above and below 10 are being compared in the same sentence, use numerals for both (e.g., “Participants read either 8 or 16 paragraphs.”). However, this sentence contains an appropriate mix of numbers and words: “Participants read eight paragraphs and then answered 20 multiple-choice questions.” The sentence is correct because the paragraphs and the questions mentioned in the sentence are different entities and so are not being compared.

When reporting a percentage, always use numerals followed by a percent sign except when beginning a sentence. This is true regardless of whether the number is less than 10 (e.g., “Only 6% of the computer games appealed

to females.”) or greater than 10 (e.g., “When using this technique, 85% of the participants improved their performance.”).

Always use numbers when describing ages (e.g., “5-year-olds”), points on a scale (e.g., “a 3 on a 5-point scale”), units of measurement (e.g., “the children stood 2 m from the target”), sample size (e.g., “6 girls and 6 boys were assigned to each study condition”), and statistics (e.g., “the mean score in the control group was 3.10”). An odd but sensible exception to the word–number rule occurs when two different types of numbers must appear together. An example is “Teachers identified fifteen 7-year-olds as the most aggressive.” This sentence avoids an awkward juxtaposition of two numbers.

For a multiplication sign, use either a lowercase *x* or the multiplication symbol used by your word processor. This is true whether you are describing a mathematical operation or a factorial design (e.g., a 2×2 design). For a minus sign, use a hyphen with a space both before and after the hyphen.

Statistical terms are abbreviated and typed with italics (e.g., *M*, *r*, *t*, *F*, *d*). When reporting the results of a statistical significance test, provide the name of the test, the degrees of freedom, the value of the test statistic, and the probability level. Here are two examples of sentences that describe statistical results:

As predicted, participants in the high-anxiety condition took longer to recognize the words ($M = 2.63$, $SD = .42$) than did the individuals in the low-anxiety condition ($M = 1.42$, $SD = .36$), $t(20) = 2.54$, $p = .02$, Cohen’s $d = 3.09$.

Job satisfaction scores were significantly correlated with marital satisfaction, $r(50) = .48$, $p < .01$.

Recall that exact probabilities (*p* values) are reported as they appear in the computer output of your statistical analysis. Use the $<$ (less than) symbol for probabilities less than .01, i.e., $p < .01$. Special symbols, including Greek letters (e.g., α) can be found using the *insert symbol* function in your word processor.

Pay attention to the way statistics are described in the articles you read. You will find that you can vary your descriptions of results to best fit your data and presentation, as well as vary your sentence constructions.

APA Style and Student Paper Formats

APA style is intended to provide a manuscript to a publisher who then prepares the paper for publication in a journal; several APA style requirements are for the convenience of the publisher. When you prepare a paper for a class report, an honours project, or a thesis, your paper may be the “final product” for your readers. When your intended audience is your instructor, you should pay close attention to what he or she has to say about expectations for the paper (Rosnow & Rosnow, 2012). In such cases, some aspects of APA style may be ignored so that your paper will closely resemble a printed report. For example, APA style calls for placement of tables and figures at the end of the paper; the publisher inserts the tables and figures in the body of the paper for the published article. However, if your report is the final version for your readers, you may need to insert tables and figures in the text or on separate pages within the body of your report. Some of the other ways that a student report may differ from APA style have been noted earlier. When you are getting ready to prepare your own report, be sure to check the particular requirements of your instructor or university.

Citing and Referencing Sources

Citing Sources in the Body of the Report

Whenever you refer to information reported by other researchers, you must accurately identify the sources. APA journals use the author–date citation method: The author name(s) and year of publication are inserted at appropriate points. The citation style depends on whether the author names are part of the narrative or are in parentheses.

One Author

When the author’s name is part of the narrative, include the publication date in parentheses immediately after the name:

Gervais (2011) found that anti-atheist prejudice can be reduced by reminding believers how prevalent atheists are in society.

When the author’s name is not part of the narrative, the name and date are cited in parentheses at the end of an introductory phrase or at the end of the sentence:

Religious believers show lower anti-atheist prejudice after being reminded atheists are prevalent in society (Gervais, 2011).

In one study (Gervais, 2011), religious believers were reminded that atheists are a prevalent group . . .

Two Authors

When the work has two authors, both names are included in each reference citation. The difference between narrative and parenthetical citations is in the use of the conjunction “and” and the ampersand “&” to connect authors’ names. When the names are part of a sentence, use the word “and”

to join the names of two authors. When the complete citation is in parentheses, use the “&” symbol:

Adults’ ability to remember to do something in the future (i.e., prospective memory) is predicted by personality traits, specifically conscientiousness (Cuttler & Graf, 2007).

Cuttler and Graf (2007) found that adults’ ability to remember to do something in the future (i.e., prospective memory) is predicted by personality traits, specifically conscientiousness.

Three to Five Authors

When a report has three to five authors, all author names are cited the first time the reference occurs. Thereafter, cite the first author’s surname followed by “and colleagues” or the abbreviation *et al.* along with the publication date. The abbreviation may be used in narrative and parenthetical citations:

First citation

Research suggests that repeating positive self-statements (e.g., “I’m a good person”) may lead people with low self-esteem to feel negatively about themselves (Wood, Perunovic, & Lee, 2009).

Wood, Perunovic, and Lee (2009) reported that people who have low self-esteem who repeated a positive self-statement (e.g., “I’m a good person”) felt worse about themselves than if they did not repeat any statement.

Subsequent citations

Among people with high self-esteem, repeating a positive self-statement has little impact on how they feel about themselves (Wood et al., 2009).

Wood et al. (2009) also examined the impact of positive self-statements on those with high self-esteem.

Another question about subsequent citations is whether to include the publication date each time an article is referenced. Within a paragraph, you do *not* need to include the year in subsequent citations as long as the study cannot be confused with other studies cited in your report.

Citation within a paragraph

In a recent study, Sana, Weston, and Cepeda (2013) . . .

Sana and colleagues also reported that . . .

Page 311 When subsequent citations are in another paragraph or in another section of the report, the publication date should be included.

Six or More Authors

Occasionally you will reference a report with six or more authors. In this case, use the abbreviation *et al.* after the first author's last name in *every* citation. Although you would not list all author names in the text, the citation in the references section of the report should include the names of the first six authors followed by *et al.* for additional authors.

References with No Author

When an article has no author (e.g., some newspaper or magazine articles), cite the first two or three words of the title in quotation marks, followed by the publication date:

Citation in text

In an article on smoking (“The decline of smoking,” 2011), data obtained from Statistics Canada . . .

Which refers readers to this citation in the reference list

The decline of smoking in Canada. (2011, July 29). CBC News.
Retrieved from <http://www.cbc.ca>

Multiple Works within the Same Parentheses

A convenient way to cite several studies on the same topic or several studies with similar findings is to reference them as a series within the same parentheses. When two or more works are by the same author(s), report them in order of year of publication, using commas to separate citations:

Mio and Willis (2003, 2005) found . . .

Past research (Mio & Willis, 2003, 2005) indicates . . .

When two or more works by different authors are cited within the same parentheses, arrange them in alphabetical order and separate citations by semicolons:

Investigations of why people procrastinate have revealed the importance of perceptions of the task. Risk factors for procrastination include optimistic predictions of how long a task will take to complete and expecting to dislike the task (Buehler, Griffin, & Ross, 2002; Buehler, Peetz, & Griffin, 2010; Steel, 2007).

Reference List Style

The APA *Publication Manual* provides examples of 97 different reference formats for journal articles, books, book chapters, technical reports, convention presentations, dissertations, Web pages, and videos, among many others. Only a few of these are presented here. When in doubt about how to construct a reference, consult the APA manual. The general format for a reference list is as follows:

1. The references are listed in alphabetical order by the first author's last name. Do not categorize references by type (i.e., books, journal articles, and so on). Note the spacing in the typing of authors' names in the examples.

2. Elements of a reference (authors' names, article title, publication date) are separated by periods.

Page 312 The first line of each reference is typed flush to the left margin; subsequent lines are indented. This is called a *hanging indent*. When you type the reference, it will appear as follows:

Stermac, L., Elgie, S., Dunlap, H., & Kelly, T. (2010). Educational experiences and achievements of war-zone immigrant students in Canada. *Vulnerable Children and Youth Studies*, 5, 97–107.

Each reference begins on a new line (think of each reference as a separate paragraph). Most word processors will allow you to easily format the paragraph with a hanging indent, so you do not have to manually insert spaces on the second and subsequent lines. Using Microsoft Word, for example, begin the paragraph with Ctrl-t (control key and t pressed simultaneously).

Format for Journal Articles

Most journals are organized by volume and year of publication (e.g., volume 66 of *American Psychologist* consists of all 12 journal issues published in 2011). A common confusion is whether to include the journal issue number in addition to the volume number. The rule is simple: If the issues in a volume are paginated consecutively throughout the volume, *do not* include the journal issue number. If each issue in a volume begins with page 1, the issue number should be included. Specific examples are shown next.

In the reference list, both the name of the journal and the volume number are italicized. Also, only the first letter of the first word in article titles is capitalized (except proper nouns and the first word after a colon or question mark). Here are some examples:

One author—no issue number

Gervais, W. M. (2011). Finding the faithless: Perceived atheist prevalence reduces anti-atheist prejudice. *Personality and Social*

Psychology Bulletin, 37, 543–556.

Two authors—use of issue number

Nguyen, T., & Trimarchi, A. (2010). Active learning in Introductory Economics: Do MyEconLab and Aplia make any difference? *International Journal for the Scholarship of Teaching and Learning*, 4(1), 1–18.

Format for Books

When a book is cited, the title of the book is italicized. Only the first word of the title is capitalized; however, proper nouns and the first word after a colon or question mark are also capitalized. The city of publication and the publishing company follow the title. If the city is not well known, include the U.S. Postal Service two-letter abbreviation for the state (e.g., AZ, NY, MN, TX); if the city is outside the United States, include the city and country (e.g., Toronto, Canada; Manchester, England).

One-author book

Carlberg, C. (2011). *Statistical analysis: Microsoft Excel 2010*. Indianapolis, IN: Que Publishing.

One-author book—second or later edition

Creswell, J. W. (2007). *Qualitative inquiry and research design: Choosing among five approaches* (2nd ed.). Thousand Oaks, CA: Sage.

Edited book

Gosling, S. D., & Johnson, J. A. (Eds.). (2010). *Advanced methods for behavioral research on the Internet*. Washington, DC: American Psychological Association.

Format for Chapters in Edited Books

For edited books, the reference begins with the names of the authors of the article, not the book. The title of the article follows. The name(s) of the book editor(s), the book title, the inclusive page numbers for the article, and the publication data for the book follow, in that order. Only the book title is italicized, and only the first letters of the article and book titles are capitalized. Here are some examples:

One editor

Brown, A. L., & Campione, J. C. (1994). Guided discovery in a community of learners. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (pp. 229–270). Cambridge, MA: MIT Press.

Two editors

Glantz, L. H., Annas, G. J., Grodin, M. A., & Mariner, W. K. (2001). Research in developing countries: Taking “benefit” seriously. In Teays, W., & Purdy, L. (Eds.), *Bioethics, justice, and health care* (pp. 261–267). Belmont, CA: Wadsworth.

Format for “Popular Articles”

The reference styles shown next should be used for articles from magazines and newspapers. As a general rule, popular press articles are used sparingly (e.g., when no scientific articles on a topic can be found or to provide an example of an event that is related to your topic).

Magazine—continuous pages

Carson, S. (2011, May/June). The unleashed mind. *Scientific American Mind*, 22(2), 22–29.

Newspaper—no author

Savannahs shaped human evolution, new study concludes. (2011, August 5). *Vancouver Sun*, p. B3.

Newspaper—discontinuous pages

Bailey, I. (2011, August 5). Euthanasia issue won't be reopened, Nicholson says. *Globe and Mail*, pp. S1, S2.

Format for Papers and Poster Sessions Presented at Conferences

Occasionally you may need to cite an unpublished paper or poster session that was presented at a professional meeting. Here are two examples:

Paper

Cheung, I., Conway, P. J., Maxwell-Smith, M., & Seligman, C. (2009, June). *Happiness and the outcome of the 2008 Canadian Federal Election*. Paper presented at the 70th annual meeting of the Canadian Psychological Association, Montreal, Canada.

Poster session

Kang, S. K., & Chasteen, A. L. (2011, January). *Beyond the double jeopardy hypothesis: The interaction between age- and race-based stereotypes across the lifespan*. Poster session presented at the meeting of the Society for Personality and Social Psychology, San Antonio, TX.

Page 314

Secondary Sources

Sometimes you need to cite an article, book, or book chapter that you read about through a textbook, an abstract, or a book review. Although it is always preferable to read and cite primary sources, sometimes you may have to cite a secondary source instead if you have exhausted all possible options for finding the source.

Suppose you wish to cite an article that you read about in a book. When you refer to the article in your paper, you need to say that it was cited in the book. In the following example, a paper by Widmeyer and McGuire is the secondary source:

Widmeyer and McGuire (as cited in Gee & Leith, 2007) suggested that playing a defensive position in hockey is associated with increased frustration and aggression . . .

In the reference list at the end of the paper, simply provide the reference for the primary source you used (in this case, the Gee and Leith citation):

Gee, C. J., & Leith, L. M. (2007). Aggressive behavior in professional ice hockey: A cross-cultural comparison of North American and European born NHL players. *Psychology of Sport and Exercise*, 8, 567–583.

Reference Formats for Electronic Sources

The amount and types of information available via the Internet has exploded. The American Psychological Association provided guidelines in the *Publication Manual* and published updates for some formats in the *APA Style Guide to Electronic Resources* (APA, 2012). There are 70 different types of references described in this guide. Only a few are provided here. The primary goal of all references is to allow readers to easily find the original source material.

Citing a Website

Sometimes you simply want to cite a particular website in your paper without referring to a specific document. In this case, just provide the address of the website. No listing in the references is necessary. For example, the following citation might appear in the text of your paper.

Most professional associations in psychology maintain websites for their members and the public. The site of the Canadian Psychological

Association is <http://www.cpa.ca> and the Association for Psychological Science site is <http://www.psychologicalscience.org>.

Citing Specific Web Documents/Pages

Many Web pages were written specifically for the Web and should not be considered as journal articles or books. For example, a document prepared by David Kenny provides information on mediating variables. In general, the rules for citing such documents are very similar to citations for journal articles. In our example, your text might read:

Kenny (2009) describes a procedure for using multiple regression to examine causal models that include mediating variables.

Page 315 Your actual reference to the document would be:

Kenny, D. A. (2009). *Mediation*. Retrieved from
<http://davidakenny.net/cm/mediate.htm>

Note that the reference includes the author, a date that was provided in the document, and a title. Some Web documents do not include a date; in this case, simply substitute n.d. in parentheses to indicate that there is no date. Most important, information is provided on the file name (URL) of the document. Note also that there is no period at the end of the reference.

There is an important rule about typing the URL (location) of the document you are citing. It is acceptable to have the URL carry over two lines if it will not fit on a single line. However, never insert a hyphen because this is not part of the address. Instead, let the address carry over with no extra hyphen.

Citing Journal Articles

Final published versions of journal articles are increasingly available via searches in a variety of library and Internet full-text databases such as *PsycINFO*. When citing these articles, the primary new feature to look for is the DOI—the digital object identifier, which is intended to help others

locate the article. The DOI is now used for research articles but not articles in the popular press. You can find the DOI as a field in databases such as *PsycINFO*; it may also appear in the text of the article, usually on the first or last page. It will appear as a long series of numbers and letters. Here is a citation that includes the DOI:

Mather, A. A., Stein, M. B., & Sareen, J. (2010). Social anxiety disorder and social fears in the Canadian military: Prevalence, comorbidity, impairment, and treatment-seeking. *Journal of Psychiatric Research*, 44, 887–893.
doi:10.1016/j.jpsychires.2010.02.01

Do not provide information on the date retrieved or URL because this is the final, published version of the paper and will not change.

Some articles that you access online will not have a DOI. In such cases, provide the standard article information. Then include “Retrieved from URL” to provide the URL of the article. There is still no need to provide the date retrieved because it is the final version of the article.

Citation of an Abstract

Sometimes you may need to cite the abstract of an article that you found in a search of *PsycINFO* or another database. Although it is preferable to find the original article, you may find that the original article is not available online or at any nearby libraries or is published in a foreign language with only the abstract available in English. Here is an example:

Morisano, D., Hirsh, J. B., Peterson, J. B., Pihl, R. O., & Shore, B. M. (2010). Setting, elaborating, and reflecting on personal goals improves academic performance. *Journal of Applied Psychology*, 95, 225–264.
Abstract retrieved from PsycINFO database.

In this example, the complete reference is given. However, you also provide the crucial information that you have only examined the abstract of the article and you found the abstract through a search of the *PsycINFO* database.

There are many other examples in the *APA Style Guide to Electronic Resources*, including books, encyclopedias, newspaper articles, and presentation slides. The rules are consistent and rely on whether the source is a final, published version; whether the DOI is provided; and whether you need to have information on the database in order to make sure you can find the document.

Conference Presentations

Students present their research findings in many different ways: in class, at regional and national meetings of professional psychology organizations, and at conferences specifically designed to highlight student research. If you are interested, ask faculty members at your institution to recommend conferences.

The presentation may take the form of a talk to an audience or a poster presentation in which other conference attendees may read the poster and engage in conversation with the presenter. Psi Chi, the International Honor Society in Psychology, has posted guidelines for paper and poster presentations on its website (www.psichi.org/?RES_ConvPresent#.Va1Rp_lVikp). We will explore the major points, but any student planning a presentation may wish to obtain more detailed advice. (As an aside, Psi Chi chapters are rare in Canada. If your university doesn't have a Psi Chi chapter, you may want to visit its website for information about how to become a member.)

Paper Presentations

Paper presentations are talks that are often only about 10 to 12 minutes long. Conference attendees typically see many of these talks during a conference. It is easy for attendees to become overloaded with information. The major thing to remember, then, is that you should attempt to convey only a few major ideas about why and how you conducted your research. Your audience wants the “big picture,” so you can avoid describing the details of past research findings, discussing exactly how you did your data analysis, or listing every step in your procedure. Omit technical jargon inappropriate for a broad audience. Use clear language to convey why the topic is important, your hypothesis, the general methods you used, and the major results. Provide a quick summary at the end, and finish with a major conclusion or key implication of the results.

The Psi Chi guidelines also advise you to write the presentation in advance but not to read it to your actual audience. You can use the written version for practice and timing. Consider bringing copies of a written summary of your presentation that includes your name, the title of the presentation, when and where it was

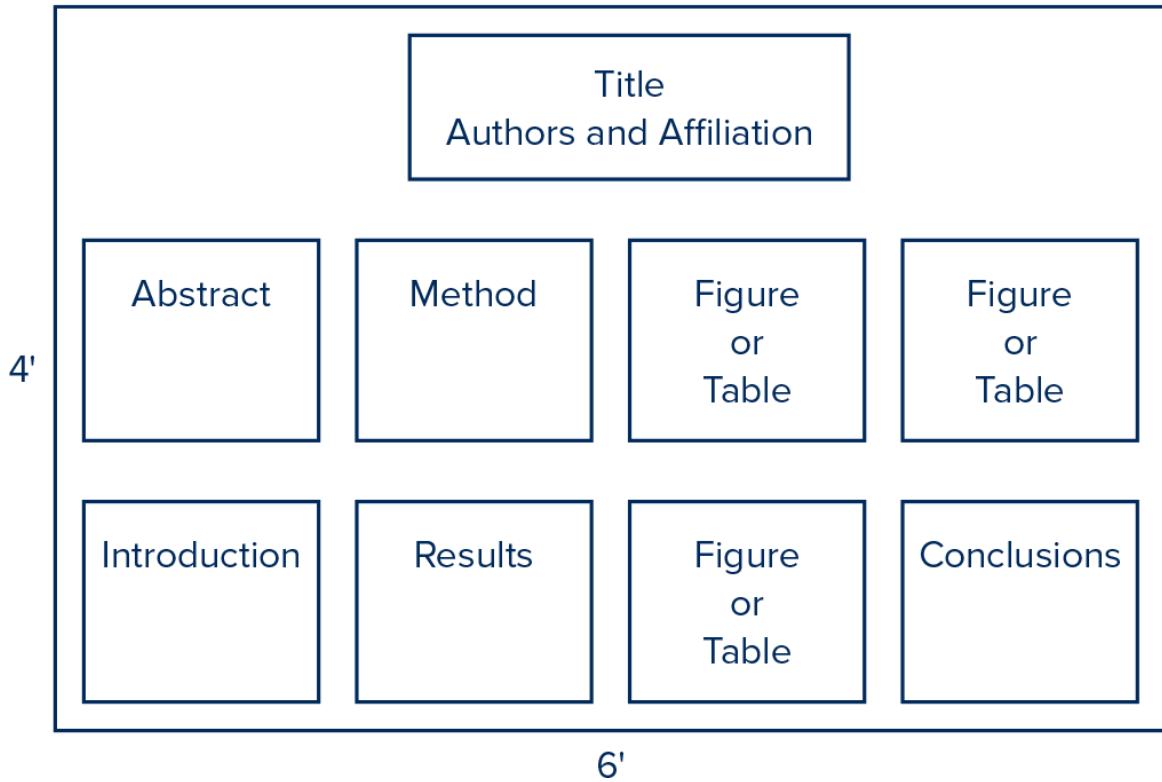
presented, and how you can be contacted. Interested audience members may wish to refer to it later.

Poster Sessions

A poster session can involve just a few or hundreds of presenters simultaneously. Each presenter is provided with space to display poster material. During the poster session, members of the audience may stop to read the poster, and some may have questions or comments. The chance to have conversations about your research with people who find your work interesting is the most valuable feature of a poster session.

Conference organizers will provide information in advance about the amount of space available for each poster. Typical dimensions are 3 to 4 feet high and 6 to 8 feet wide. The content of each poster will usually be divided up like an APA style manuscript, although it will contain much less information for each section. An example poster layout is provided in Figure A.1. The Psi Chi website has other suggested layouts. The actual construction of the poster may consist of a series of separate pages or a single professionally printed poster using large format printing technology.

FIGURE A.1 A sample poster



Avoid providing too much detail—often a bulleted list of major points will be most effective. One or two easy-to-read figures can also be very helpful. There are probably no more than two major points that you would like someone to remember after viewing your poster. Make sure those points are obvious. The font that you use should be large enough to be read from a distance (usually the text will be 18-point font). Colour can be used to enhance the attractiveness of the poster. For interested readers, bring copies of a summary poster handout that includes your contact information as well as the date and location of the conference.

Appendix A Sample Paper

The remainder of this appendix consists of a typed manuscript of a paper that was published in a professional journal. This is intended to be a useful guide when you write your own reports in APA style. Read through the manuscript, paying particular attention to the general format, and make sure you notice the rules concerning page numbering, section headings, reference citation, and the format of figures. Writing your first research report is always a challenging task. It will become easier as you read the research of others and gain practice by writing reports of your own.

Janet Polivy, C. Peter Herman, and Rajbir Deo graciously gave their permission to reprint their paper to illustrate elements of APA style. The comments at the side alert you to features of APA style that you will need to know about when writing your own papers. Be aware, though, that every paper will include slightly different types of information depending on the particular topic, method, and results. Your paper will follow the general guidelines of APA style, but many of the details will be determined by the needs of your study.

You may note that the title of this paper is longer than the 10 to 12 words recommended by APA style. A briefer (but less intriguing) title might omit the part before the colon. Also note that the horizontal lines are meant to signal page breaks. Do not include these lines in your manuscript (except in a Figure).Page 318

1

Running head: Getting a Bigger Slice

Getting a Bigger Slice of the Pie: Effects on Eating and Emotion in Restrained and Unrestrained *Eaters*

Janet Polivy, C. Peter Herman, and Rajbir *Deo*

University of Toronto

Page 319

2

Getting a Bigger Slice

Abstract

We investigated the influence of perceptions of the portion size of food on subsequent eating by restrained and unrestrained eaters. In the present study, all participants were served a same-sized slice of pizza. For one-third of participants, their slice appeared larger than the slice being served to another ostensible participant, another third perceived their slice as smaller, and the final third did not see a second slice. All participants then proceeded to “taste and rate” cookies in an ad lib eating opportunity. A significant interaction reflected the fact that when the pizza slice was perceived as large, restrained eaters tended to eat more cookies whereas unrestrained eaters tended to eat less cookies. Emotion data suggest that the differential responses of restrained and unrestrained eaters to the belief that they have overeaten relative to another eater influenced their subsequent dissimilar ad lib eating behavior.

Page 320
3

Getting a Bigger Slice

Getting a Bigger Slice of the Pie: Effects on Eating and Emotion in Restrained and Unrestrained Eaters

We often eat one food followed by another (e.g., main course and then dessert). How much we eat of the later food probably depends to a large extent on our intake of the earlier food. In the laboratory, we refer to the earlier food as a “preload.” The effects of food preloads on subsequent eating are complex: Chronic dieters or restrained eaters generally respond quite differently than nondieters or unrestrained eaters do. Whereas unrestrained eaters typically compensate by eating less after a larger preload than after a smaller one, restrained eaters often “counter-regulate,” eating more after a large preload than after a small preload or after no preload at all ([Adams & Leary, 2007](#); Herman, Polivy, & [Esses, 1987](#); [Polivy, Heatherton, & Herman, 1988](#); Polivy, Herman, Hackett, & Kuleshnyk, 1986). Presumably, the larger preload is more likely to sabotage the restrained eater’s diet for that day, undermining motivation for continued restraint and unleashing disinhibited eating (possibly potentiated by chronic perceived deprivation). If the preload is actually large and fattening, it is likely to produce disinhibited eating by restrained eaters ([Herman et al., 1987](#); McCann, Perri, Nezu, & Lowe, 1992; Polivy et al., 1986, 1988), but disinhibition may be observed even when the restrained eater is merely led to believe that the

preload is high in calories or otherwise forbidden (Polivy, 1976; Spencer & Fremouw, 1979) or when the restrained eater draws that implication from the nature of the food itself (Knight & Boland, 1989). Previous studies have manipulated the perceived size of the preload (holding actual size or caloric content constant) by either telling participants that the preloads are high in calories (*Polivy, 1976*; Spencer & Fremouw, 1979) or by implication. For example, Knight and Boland (1989), served iso-caloric preloads of milkshake or a cottage cheese and fruit mixture. Restrained eaters displayed disinhibition only when served milkshake, because they regard milkshake as inherently more caloric than the salad-like cottage cheese and fruit mixture. More recently, *Pliner and Zec (2007)* showed that thinking about a preload as a meal (rather than as a snack) makes people perceive the preload as higher in calories and affects eating accordingly.

The present study was designed to further extend the exploration of the effects of perceived preload/portion size on eating and, moreover, to do this in a more externally valid meal setting. In order to understand the source of these effects, we included measures of affective responses, as it has been shown that affect influences eating differently for restrained and unrestrained eaters (Polivy & Herman, 1999), and eating, especially eating what is seen as a large amount, affects emotions differently in restrained and unrestrained eaters (Polivy & Herman, 2005).

We hypothesized that even with the preload/meal held constant, restrained eaters who regard the portion as larger will subsequently eat more than will those who regard it as normal sized or small, because the “large” portion is more likely to break their diets and lead to disinhibited eating. We also predicted that unrestrained eaters will eat less after a perceived large portion than after what they perceive to be a normal-sized meal, and less after a portion perceived as normal sized than after one perceived as small. In the present study, all participants received an identical, standard light lunch meal, but some were led to perceive the portion they received as large and some were led to perceive the portion as small simply by means of social comparison (or more accurately, perceptual contrast). If someone gets a larger portion than yours, your own portion may appear to be “small,” whereas, if someone gets a portion that is smaller than yours, your own portion may appear to be “large.”Page 321

Method

Participants

The participants were [106](#) female undergraduate students enrolled in an introductory psychology class at a large university. The participants were recruited via an experimental database, where they could sign up for a study entitled “Market Taste Test Study.” Each experimental session lasted [1 hr](#); participants received credit toward a course requirement for their participation.

Materials

Food. Extra-large uncut cheese pizza was ordered from a local pizza chain for each day of experimentation. The slices were reheated in a microwave oven before they were served to the participants. Any leftover pizza was stored in a freezer for use on another day. Frozen cookie dough (from a manufacturer who supplies “fresh-baked” cookies to local restaurants) was stored in a freezer and used to bake bite-sized cookies regularly throughout the week. Three different types of cookies were baked as needed: oatmeal raisin, chocolate chip, and double chocolate chip.

Questionnaires. Pre- and post-pizza rating scales were completed by all participants immediately after the manipulation and again after the pizza but before the cookie taste test. The pizza rating scales included a section from the [PANAS](#) (Positive and Negative Affect Schedule; Watson, Clark, & Tellegen, 1988), which was designed to measure the participants’ negative affect. Participants described their negative emotional states such as “guilty” and “angry” using rating scales ranging from 1 *for very slightly or not at all* to 5 for *extremely*. Other questions assessed hunger and various aspects of the pizza that they were about to eat or had just eaten using a 9-point Likert-type scale. These questions included a manipulation check that asked participants to rate the portion size (rating from 1 for *too small* to 5 for *just right* to 9 for *too big*). The questions were answered before and after eating the pizza (with wording changed appropriately). In addition, participants were asked at the end of the study to what extent they had noticed any difference in the size of the pizza slices (“How did your slice of pizza compare to the slice received by the other person in the study?” Response options were *smaller*, *the same*, and *larger*).Page 322

Restraint scale. The 10-item Herman and Polivy Revised Restraint Scale (Herman, Polivy, & Silver, 1979) was used to determine restraint status. Participants who scored 15 or less on the restraint scale were classified as unrestrained eaters, whereas participants who scored above 15 were classified as restrained eaters.

Procedure

Female participants were recruited for this study through a psychology experimental website advertisement that specified that the participants must have no food allergies, must not be lactose intolerant, and should refrain from eating for up to 3 hr prior to their experimental session.

Each participant was informed that she would be given a light vegetarian cheese pizza lunch in order to ensure that each participant had the same taste experience and same level of fullness before completing taste ratings for market research. She was told that she would be sampling various food products that were being proposed for the market by a large food company that targeted the university-student population. The participant was also informed that she would be completing some questionnaires to assess her mood and other variables to ensure that these factors were not influencing her food ratings. Furthermore, the participant was told that she would be discussing her food ratings with another female participant in a brief discussion at the end of their session. She then signed the consent form.

Participants were randomly assigned to one of three pizza-slice conditions: smaller slice, larger slice, and no information. Regardless of which condition the participant was in, she always received a standard slice of pizza (1/6 of the pizza), but the size of the “other participant’s” slice was varied. Each pizza was cut into six pieces consisting of four standard-sized slices (1/6 of the pizza), one larger slice (1/3 larger than a standard-sized slice), and one smaller slice (1/3 smaller than the standard slice). In order to ensure that each slice was consistently cut for all pizzas used in the study, the appropriate sized slices were drawn onto a piece of paper and cut out to be used as templates for all pizzas. Thus, in the “smaller” condition, the participant received a standard-sized slice of pizza, while the “other female participant” was supposedly receiving the slice 1/3 larger than the standard slice. Similarly, participants in the “larger” condition received the standard slice of pizza, while the “other female participant” appeared to be receiving the slice 1/3 smaller than the standard slice. In the “no information”

control condition, participants were given a standard-sized slice of pizza, with no indication of the “other female participant’s” slice. Page 323

6

Getting a Bigger Slice

When the experimenter presented the participant with the pizza slices, the pizza slices were placed on a tray with a glass of water next to each slice and brought into the experimental room. Each participant in the “smaller” condition was presented with her standard-sized slice of pizza next to the 1/3 larger slice belonging to the “other female participant,” which was identified as such as it was situated further away from her. Each participant in the “larger condition” saw her standard-sized slice along with the “other” female participant’s smaller slice. In the “no information” control condition, the participant was presented only with her standard-sized slice, along with a glass of water. The experimenter then left the room, leaving the slices in the room and explaining that she had to retrieve a questionnaire for the participants. The experimenter left the room for exactly 1 min, allowing a sufficient amount of time for the participant to observe the slices and perceive the differences in their sizes. When the experimenter returned, the participant’s slice and water were placed on the table in front of her and the pre-pizza rating scale was handed to the participant. She was asked to complete the questionnaire before eating her pizza slice. The experimenter then left the room with the “other participant’s” slice.

The participant was given 7 min to complete the preeating questionnaire and to eat her entire pizza slice (supposedly to ensure equal fullness in all participants), after which time the experimenter returned to the experimental room and handed the participant another set of questionnaires (to maintain the cover story). These questionnaires included the post-pizza scales. The participant was instructed to ring a bell when she had completed the questionnaires. At that time, the experimenter returned with three heaping (preweighed) plates of each of three types of cookie and another glass of water, plus three cookie-rating sheets (one for each cookie type). Tasting these cookies was ostensibly the principal purpose of the experiment, but the cookies were actually provided as a measure of ad lib consumption. In order to measure how many cookies the participants ate, the cookies were weighed prior to the experimental session and again after the “taste-test task.” A heaping amount of each cookie type (oatmeal raisin, chocolate chip, and double chocolate chip) was placed onto one of three separate plates and the weight of each plate was measured and recorded. The three plates of cookies

were placed on the table in front of the participant, with oatmeal-raisin cookies always being first, chocolate-chip cookies second, and double chocolate-chip cookies third. The participant was instructed that she would now be participating in the taste-test portion of the study, wherein she would sample three different types of cookies that were about to be released on the market by a large food company that marketed its snack foods to the university-aged population. The participant was instructed to begin with the oatmeal-raisin cookies and take as many cookies as she required to be very sure of her taste ratings of the cookies. She was told to sample the oatmeal-raisin cookies first, followed by the chocolate chip cookies, and finally the double chocolate-chip cookies. It was emphasized that once she had completed the ratings for one cookie type, she was not to go back and resample the previous cookie type and she was not to change her ratings once she had moved on to a “new taste.” The water was provided to permit the participant to “cleanse her palate” as she moved from cookie to cookie. Moreover, the participant was reminded to be sure of her ratings since she would be comparing her food ratings with the “other female participant” at the end of the session. Finally, the participant was informed that once she was finished making her ratings, there were plenty of cookies and she was free to have as many more of any type as she liked, as long as she did not change any ratings. After the instructions were clear to the participant, the experimenter left the room for 10 min.

Page 324

7

Getting a Bigger Slice

The experimenter reentered the room with the final set of questionnaires to be completed (including the restraint scale). The cookie plates were removed from the room, where they were reweighed without the participants’ knowledge, in order to measure how many grams of cookies the participant had consumed. When the participant had completed the last set of questionnaires, her height and weight were measured and recorded. The participant was debriefed as to the purpose of the study. She was also asked if she had noticed the size difference in the pizza slices that had been presented, when she last ate, and what she ate at that time. She was thanked and asked some questions about the experiment so that she could receive credit for her psychology course before being dismissed. The study was thus conducted in accordance with ethical principles and had full institutional ethical review and approval.

Results

Participant Characteristics

A series of 2 (restrained versus unrestrained) \times 3 (control, “small slice,” “large slice”) ANOVAs indicated the usual restraint main effect on BMI, $F(1,98) = 9.77, p = .002$, with restrained eaters having higher BMIs ($M = 24.27$) than unrestrained eaters ($M = 21.63$). There was no effect of condition and no significant interaction; as well, there were no restraint or condition differences in preeating hunger.

Manipulation Checks

On the pre- and post-pizza questionnaires, participants were asked about the quantity of pizza that they had been served. A 2 (restrained versus unrestrained) \times 3 (control, “small slice,” “large slice”) ANOVA on each of these questions yielded only main effects for condition, preeating $F(2, 97) = 5.25, p = .008$, and posteating $F(2, 100) = 8.16, p < .001$. In both analyses, the “small” slice was rated as close to 5 (which corresponded to “just right”) ($M_{pre} = 5.22; M_{post} = 5.43$), the control/no information slice was seen as bigger than the small one ($M_{pre} = 5.69; M_{post} = 6.03$), and the “large” slice was seen as bigger than either of the others ($M_{pre} = 6.33; M_{post} = 6.80$). All differences were significant at the .05 level.
Page 325

8

Getting a Bigger Slice

Cookie Intake

A 2 (Restraint: restrained, unrestrained) \times 3 (Condition: control, “small slice,” “large slice”) ANOVA on the amount of cookies eaten (in grams) yielded no main effects of restraint or condition; however, there was a significant interaction, $F(2, 100) = 3.51, p = .034, \eta^2 = 0.066$. Post hoc t tests indicated that whereas neither restrained nor unrestrained participants in the “small slice” condition differed from those in the control condition or from each other, restrained and unrestrained eaters in the “large slice” condition differed significantly from each other, $t(100) = 2.98, p = .005$. In addition, although not significant, restrained eaters in the “large slice” condition ate marginally more than did restrained eaters in the control condition, $t(100) = 1.82, p = .075$, and unrestrained eaters in the “large slice” condition ate marginally less than did unrestrained eaters in the

control condition, $t(100) = 1.66$, $p = .10$. (see Table [1](#) for all means and standard deviations).

Negative Affect

A 2×3 ANOVA on total negative affect before eating the pizza (but after the manipulation of perceived portion size) yielded no significant main effects, but there was a significant interaction between restraint and condition, $F(2, 100) = 3.40$, $p = .037$, $\eta^2 = 0.066$ (see [Figure 1](#)). The only significant differences found in the post hoc t tests were between the “small” versus “large” conditions for the unrestrained eaters, with those receiving the large slice feeling more negative emotion than those receiving the small slice, $t(100) = 2.25$, $p = .026$, and between restrained and unrestrained eaters in the “small” condition, $t(100) = 2.03$, $p = .045$, with restrained eaters feeling more negative affect than did unrestrained eaters. The analysis comparing restrained and unrestrained participants in the “large” slice condition indicated a trend in the opposite direction, $t(100) = 1.49$, $p = .14$, as did the analysis comparing “small” versus “large” for restrained eaters, $t(100) = 1.41$, $p = .17$. The negative affect ratings made after eating the pizza were no longer significantly different. Also, there were no significant effects on hunger ratings either before or after eating the pizza.

[Discussion](#)

Participants clearly perceived the size of their portion of pizza differently as a function of whether they saw a comparison slice and what they saw in the comparison. When they saw their slice next to a larger one, they perceived their slice as smaller than did those who did not see a comparison slice; and when they saw their slice next to a smaller slice they perceived it as larger than did participants who saw only their own slice. The change in perception occurred despite the fact that not only were all participants given the same-sized slice, but this size is the standard slice sold on campus and at all other outlets of the major pizza chain that supplied the pizza.

Getting a Bigger Slice

Based simply on these different perceptions of the identical portion, participants went on to eat different amounts, as shown by the significant interaction between restraint and condition. Those who saw their pizza slice as smaller ate the same

amount of cookies as did those who did not have a comparison (regardless of restraint status), but the cookie intake of those who thought that they had eaten a larger slice of pizza was affected by this perception (in different ways depending on restraint status). That the effect was more a matter of the “large slice” condition changing intake somewhat (relative to control) than of the “small slice” condition changing intake (relative to control) was probably due to the fact that in the “small slice” condition, the slice was seen as close to—indeed, slightly more than—“just right,” but the “large slice” was seen as significantly larger than “just right.” In other words, it was the pizza in the “large slice” condition that was seen as unusually large, rather than the pizza in the “small slice” condition being seen as unusually small.

The direction of the effect in the “large slice” condition depended on restraint status, as was reflected in the significant interaction. Restrained and unrestrained eaters ate the same amount of cookies in the control and “small slice” conditions, but unrestrained eaters tended to eat less in the “large slice” condition, whereas restrained eaters tended to eat more in the “large slice” condition. (Restrained eaters ate significantly more than did unrestrained eaters in the “large slice” condition.) In other words, unrestrained eaters compensated by eating less cookies if they thought that they had eaten a lot of pizza, whereas restrained eaters counter-regulated and ate more cookies when they thought that they had already overeaten on pizza. This pattern corresponds to the effect obtained in previous research when preload size was actually manipulated or when perceived preload size was manipulated by telling participants that the preloads varied in caloric value (Polivy, 1976) or by implying that the preloads differed in caloric value because they were either “forbidden” or “allowed” foods (e.g., Knight & Boland, 1989). While subtle manipulations such as the smell of food have been shown to induce restrained eaters to eat more (Fedoroff, Polivy, & Herman, 1997, 2003; Jansen & Van den Hout, 1991), the present study *involved* arguably the subtlest manipulation yet. Nothing was said about size or caloric value of the preload, and the identical preload/meal was used in all conditions; only a visual comparison to someone else’s smaller portion acted to render one’s own portion relatively large, with predictable effects on subsequent intake. Moreover, the present “preload” was actually a meal (“light lunch”) rather than extraneous eating. However, even simply perceiving one’s meal as “larger than just right” seems to have been enough to push eating in opposite directions for restrained versus unrestrained eaters. Eating was thus more strongly influenced by social comparison and the perception this fostered (I’m eating more than she is) than by actual portion size.

Page 327

Getting a Bigger Slice

The fact that restrained eaters ate somewhat more in the “large slice” condition is what we have come to expect from the literature in which restrained eaters typically overeat after a large preload (or a preload perceived as large). That they did not eat less in the “small slice” condition than in the control condition was consistent with the finding that restrained eaters’ eating is “dichotomous”: they ate either a small, reasonable amount, when they were not disinhibited (i.e., when the preload was not seen as large, or they were not disinhibited by food cues, negative emotion, or other factors) or they ate a large amount when they became disinhibited (i.e., when the preload—or in the present case, meal—was large or perceived as large). In the present study, the “large slice” condition was perceived as a large preload/meal whereas the other two conditions were seen as appropriate sized.

The unrestrained eaters on the other hand did not compensate for receiving the smaller piece of pizza by eating more cookies, even though they perceived the smaller portion as smaller than the other portions. They did, however, rate the small slice as “just right” in size. They may possibly have simply responded to their internal signals of satiety and thus ate the same amount of cookies as did the unrestrained eaters who got no comparative information (and also saw their slice as close to the “right” size). Of course, unrestrained eaters in the large slice condition presumably had the same satiety signals, but unrestrained eaters may be more prepared to eat less (after a preload/meal that they consider to be large) than to eat more (after a preload/meal that they consider to be “just right”).

Not surprisingly, because all participants were actually given and ate the same amount of pizza, there were no group or condition differences in hunger either before or after eating the pizza. There were, however, some potentially interesting differences in the extent of negative affect experienced upon realizing that one had been given a larger or smaller portion of pizza. For unrestrained eaters, getting a large slice made them more dysphoric, but for restrained eaters, dysphoria was higher when they received a smaller slice. Although small, these opposite emotional reactions may speak to the differential psychology of the restrained and unrestrained eaters. Unrestrained eaters may be responding to the prescriptive norm of not appearing to eat excessively (Herman, Roth, & Polivy, 2003), and feel worse if they think that they are violating the norm. Restrained eaters, on the other hand, may actually be more upset with being allowed to

maintain their diets (by eating the smaller piece); apparently they feel somewhat better when “forced” by the experimenter to eat “more,” break their diets, and indulge themselves with additional cookies. This interpretation comports with the assumption that fundamentally, people want to eat as much as possible, but are constrained by considerations of social propriety (not eating excessively so as not to look like a “pig”) or their self-imposed dietary agendas (Herman et al., 2003). When forced by someone else to transgress against their diets, restrained eaters may well experience what we have called the “what the hell effect” (Herman & Polivy, 1984) and feel relieved to be pushed off their diets and allowed to unleash their eating. Page 328

11

Getting a Bigger Slice

The present study shows that the mere perception that one’s meal was excessively large acts the same as a gratuitous preload to disinhibit eating in restrained eaters. The data also show that restrained and unrestrained eaters alike judge the amount that they are served in comparison to what those around them are eating. Such perceptions about the social context or meaning of one’s portion apparently outweigh feelings of hunger in influencing the amount eaten, particularly if one sees oneself as having overeaten relative to others. Restrained eaters, when they perceive themselves as having eaten excessively compared to others, continue to eat liberally rather than curtail their intake. This indulgence undermines their stated dietary goals, but the fact that they feel worse when they do not (get to) overindulge provides a hint as to why dieters so often find themselves breaking their diets. Page 329

12

Getting a Bigger Slice

References

- Adams, C., & Leary, M. (2007). Promoting self-compassionate attitudes toward eating among restrictive and guilty eaters. *Journal of Social & Clinical Psychology*, 26, 1120–1144. [doi:10.1521/jscp.2007.26.10.1120](https://doi.org/10.1521/jscp.2007.26.10.1120)
- Fedoroff, I., Polivy, J., & Herman, C. P. (1997). The effect of pre-exposure to food cues on the eating behavior of restrained and unrestrained eaters. *Appetite*, 28, 33–47. doi:10.1006/appc.1996.0057

- Fedoroff, I., Polivy, J., & Herman, C. P. (2003). The specificity of restrained versus unrestrained eaters' responses to food cues: General desire to eat, or craving for the cued food? *Appetite*, 41, 7–[13](#). doi:10.1016/S0195-6663(03)00026-6
- [Jansen, A.](#), & Van den Hout, M. (1991). On being led into temptation: “Counterregulation” of dieters after smelling a preload. *Addictive Behaviors*, 16, 247–253. doi:10.1016/0306-4603(91)90017-C
- Herman, C. P., & Polivy, J. (1984). A boundary model for the regulation of eating. *Psychiatric Annals*, 13, 918–927.
- Herman, C. P., Polivy, J., & Silver, R. (1979). Effects of an observer on eating behavior: The induction of “sensible” eating. *Journal of Personality*, 47, 85–99. doi:10.1111/j.1467-6494.1979.tb00616.x
- Herman, C. P., Polivy, J., & Esses, V. M. (1987). The illusion of counter-regulation. *Appetite*, 9, 161–169. doi:10.1016/S0195-6663(87)80010-7
- Herman, C. P., Roth, D., & Polivy, J. (2003). Effects of the presence of others on food intake. A normative interpretation. *Psychological Bulletin*, 129, 873–886. doi:10.1037/0033-2909.129.6.873
- Knight, L., & Boland, F. (1989). Restrained eating. An experimental disentanglement of the disinhibiting variables of calories and food type. *Journal of Abnormal Psychology*, 98, 412–420. doi:10.1037/0021-843X.98.4.412
- McCann, K. L., Perri, M. G., Nezu, A. M., & Lowe, M. R. (1992). An investigation of counterregulatory eating in obese clinic attenders. *International Journal of Eating Disorders*, 12, 161–169. doi:10.1002/1098-108X(199209)12:2<161::AID-EAT2260120206>3.0.CO;2-A
- Pliner, P., & Zec, D. (2007). Meal schemas during a preload decrease subsequent eating. *Appetite*, 48, 278–288. doi:10.1016/j.appet.2006.04.009
- Polivy, J. (1976). Perception of calories and regulation of intake in restrained and unrestrained subjects. *Addictive Behaviors*, 1, 237–243. doi:10.1016/0306-4603(76)90016-2

Getting a Bigger Slice

- Polivy, J., Heatherton, T. F., & Herman, C. P. (1988). Self-esteem, restraint, and eating behavior. *Journal of Abnormal Psychology*, 97, 354–356. doi:10.1037/0021-843X.97.3.354
- Polivy, J., & Herman, C. P. (1999). Distress and *eating*: Why do dieters overeat? *International Journal of Eating Disorders*, 26, 153–164. doi:10.1002/(SICI)1098-108X(199909)26:2<153::AID-EAT4>3.0.CO;2-R
- Polivy, J., & Herman, C. P. (2005). Mental health and eating behaviours: A bi-directional relation. *Canadian Journal of Public Health*, 96, 43–48.
- Polivy, J., Herman, C. P., Hackett, R., & Kuleshnyk, I. (1986). The effects of self-attention and public attention on eating in restrained and unrestrained subjects. *Journal of Personality and Social Psychology*, 50, 1203–1224. doi:10.1037/0022-3514.50.6.1253
- Spencer, J. A., & Fremouw, W. J. (1979). Binge eating as a function of restraint and weight classification. *Journal of Abnormal Psychology*, 88, 262–267. doi:10.1037/0021-843X.88.3.262
- Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scale. *Journal of Personality and Social Psychology*, 54, 1063–1070. doi:10.1037/0022-3514.54.6.1063

Page 331

Getting a Bigger Slice

Table 1

Amount of Cookies Eaten (in Grams) in the Pizza Size Conditions

<u>Larger slice</u>	Control/no info	Smaller slice
---------------------	-----------------	---------------

Restraint	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>
Unrestrained eaters	50.39	6.83	25	67.89	8.04	18	61.20	7.12	23
Restrained eaters	84.36	9.13	<u>14</u>	59.94	9.86	12	62.12	9.13	14

Note. The number of participants (*n*) in the unrestrained eaters group is higher than the number in the restrained group when using the standard cutoff score on the Restraint Scale.

Page 332

15

Getting a Bigger Slice

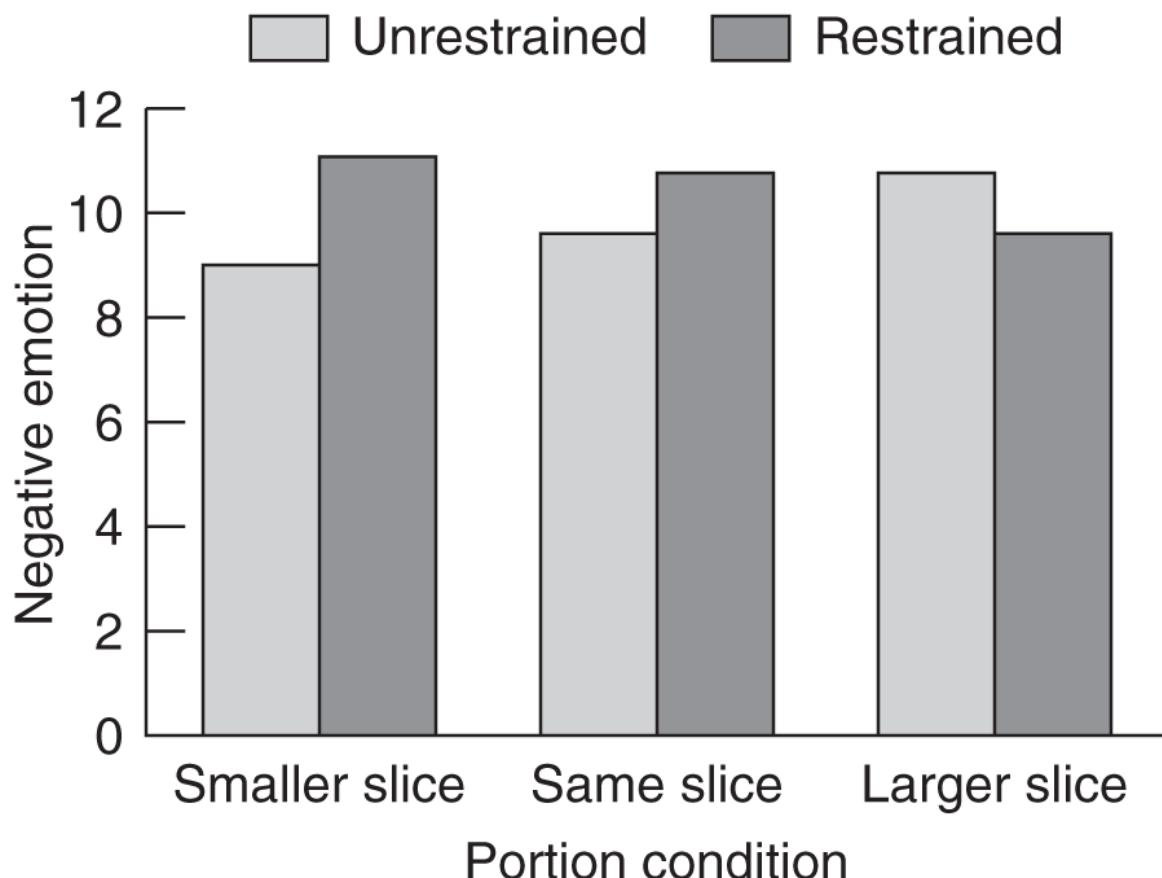


Figure 1. Total negative affect before eating pizza (but after seeing it). Post hoc *t* tests show a significant difference between the restrained and unrestrained eaters who were given the smaller slice. Unrestrained eaters getting the small slice felt better than those given the larger slice;

restrained eaters given the small slice felt marginally worse than those given the large slice.



Statistical Tests

The purpose of this appendix is to provide some formulas and calculation procedures for some data analysis. Not all possible statistical tests are included, but a variety of tests are given that should be appropriate for many research designs you might use.

We will build on [Chapters 12](#) and [13](#) to examine procedures for both descriptive and inferential statistics. You may recall from those chapters that the appropriate statistical analysis is determined by the type of design and by the measurement scale used in the study. Remember that there are four types of measurement scales: nominal, ordinal, interval, and ratio. Nominal scales have no numerical properties, ordinal scales provide rank-order information only, and interval and ratio scales have equal intervals between the points on the scale. In addition, ratio scales have a true zero point. As we consider each statistical test, we will note the relevant measurement scale restrictions that apply.

The examples here use small and simple datasets, so the calculations can be done easily by hand using a calculator. You will probably use a computer program such as SPSS, R, or Excel for your analyses. Practising the underlying calculations will help you understand the output from these computer programs.

Descriptive Statistics

Measures of Central Tendency

A measure of central tendency gives a single number that describes how an entire group scores as a whole or on the average. Three different central tendency measures are available: the mode, the median, and the mean. [Table B.1](#) shows a set of scores alongside the descriptive statistics.

Table B.1 Descriptive Statistics for a Set of Scores

Score	Descriptive Statistic
1	Mode = 5
2	
4	Median = 5
4	
5	$X = \frac{\sum X}{N} = 4.5$
5	
5	
6	Range = 6
6	
7	$s^2 = \frac{\sum (X - \bar{X})^2}{N - 1} = \frac{\sum X^2 - NX^2}{N - 1} = \frac{233 - 202.5}{9} = 3.388$
$\sum X = 45$	
$\sum X^2 = 233$	
$N = 10$	
$s = \sqrt{s^2} = 1.84$	

The Mode

The mode is the most frequently occurring score. The most frequently occurring score in [Table B.1](#) data is 5. No calculations are necessary to find the mode. The mode can be used with any of the four types of measurement scales. However, with nominal scale data, mode is the *only* measure of central tendency that can be used. If you are measuring sex of participants and find there are 100 females and 50 males, the mode is “female” because this is the most frequently occurring category on the nominal scale.

The Median

The median is the score that divides the group in half: 50 percent of the scores are below the median and 50 percent are above the median. When the scores have been ordered from lowest to highest (as in [Table B.1](#)), the median is easily found. If there are an odd number of scores, you simply find the middle score. (For example, if there are eleven scores, the sixth score is the median, because there are five lower and five higher scores.) If there is an even number of scores, the median is the average of the two middle scores. In the data in [Table B.1](#), there are ten scores, so the fifth and sixth scores are the two middle scores. To find the median, we add the two middle scores and divide by 2. Thus, in [Table B.1](#), the median is [Page 335](#)

$$\frac{5 + 5}{2} = 5$$

The median can be used with ordinal, interval, or ratio scale data. It is most likely to be used with ordinal data, however. This is because calculation of the median considers only the rank ordering of scores and not the actual

size of the scores.

The Mean

The mean does take into account the actual size of the scores. Thus, the mean is based on more information about the scores than either the mode or the median. However, it is appropriate only for interval or ratio scale data. The mean is the sum of the scores in a group divided by the number of scores. The calculational formula for the mean can be expressed as

$$X = \frac{\sum X}{N}$$

where X is the symbol for the mean. In this formula, X represents one person's score, and the \sum symbol indicates addition. The symbol $\sum X$ can be read as "sum of the X s," which indicates that everyone's scores are to be added. Thus, $\sum X$ in the data from [Table B.1](#) is

$$1 + 2 + 4 + 4 + 5 + 5 + 5 + 6 + 6 + 7 = 45$$

The N in the formula symbolizes the number of scores in the group. In our example, $N = 10$. Thus, we can calculate the mean:

$$X = \frac{\sum X}{N} = \frac{45}{10} = 4.5$$

Page 336

Measures of Variability

In addition to describing the central tendency of the set of scores, we want to describe how much the scores vary among themselves. How much spread is there in the set of scores?

The Range

The range is typically calculated as the highest score minus the lowest score, although some variations account for rounding. In our example, the range is 6. The range is not a very useful statistic, because it is based on only two scores in the distribution. It ignores all of the information about dispersion that is available across the entire set of scores.

The Variance and Standard Deviation

The variance and a related statistic called the standard deviation use all the scores to yield a measure of variability. The variance indicates the degree to which scores vary about the group mean. The formula for the variance (symbolized as s^2) is

$$s^2 = \frac{\sum (X - \bar{X})^2}{N - 1}$$

where $(X - \bar{X})^2$ is an individual score, X , minus the mean, \bar{X} , and then squared. Thus $(X - \bar{X})^2$ is the squared deviation of each score from the mean. The Σ sign indicates that these squared deviation scores are to be added together. Finally, dividing by $N - 1$ gives the mean of the squared deviations. The variance, then, is the mean of the squared deviations from the group mean. (Squared deviations are used because simple deviations would add up to zero. $N - 1$ is used in most cases for statistical purposes because the scores represent a sample and not an entire population. As the sample size becomes larger, it makes little difference whether N or $N - 1$ is used.)

The data in [Table B.1](#) can be used to illustrate calculation of the variance. $\sum (X - \bar{X})^2$ equals

$$(1 - 4.5)^2 + (2 - 4.5)^2 + (4 - 4.5)^2 + (4 - 4.5)^2 + (5 - 4.5)^2 + (5 - 4.5)^2 + (5 - 4.5)^2$$

The next step is to divide $\sum (X - \bar{X})^2$ by $N - 1$. The calculation for the variance, then, is

$$s^2 = \frac{\sum (X - \bar{X})^2}{N - 1} = \frac{30.50}{9} = 3.388$$

You might encounter a different calculational formula for the variance:

$$s^2 = \frac{\sum X^2 - \bar{X}^2}{N - 1}$$

where $\sum X^2$ is the sum of the squared individual scores, and \bar{X}^2 is the mean squared. You can confirm that the two formulas are identical by computing the variance using this simpler formula. (Remember that $\sum X^2$ tells you to square each score and then add the squared scores.)

The standard deviation is the square root of the variance. Because the variance uses squared scores, the variance does not describe the amount of variability in the same units of measurement as the original scale. The standard deviation (s) corrects this problem. Thus, the standard deviation is the average deviation of scores from the mean.

Pearson Product-Moment Correlation Coefficient

As explained in [Chapters 4](#) and [12](#), the Pearson correlation coefficient (r) is used to describe the strength of the relationship between two variables that have been measured on interval or ratio scales.[Page 337](#)

Example

Suppose you want to know whether travel experiences are related to knowledge of geography. In your study, you give a 15-item quiz on North American geography, and you also ask how many American states and Canadian provinces participants have visited. After obtaining the pairs of observations from each participant, a Pearson r can be computed to measure the strength of the relationship between travel experience and knowledge of geography.

[Table B.2](#) presents fictitious data from such a study along with the calculations for r . A formula to calculate r is

$$r = \frac{N\sum XY - \sum X/\sum Y}{\sqrt{N\sum X^2 - (\sum X)^2} \sqrt{N\sum Y^2 - (\sum Y)^2}}$$

where X refers to a participant's score on variable X , and Y is a participant's score on variable Y . In [Table B.2](#), the travel experience score is variable X , and the geography knowledge score is variable Y . In the formula, N is the number of paired observations (that is, the number of participants measured on both variables).

Table B.2 Data for Hypothetical Study on Travel and Knowledge of Geography:
Pearson r

Subject Identification Number	Travel Score (X)	Knowledge Score (Y)	XY
01	4	10	40
02	6	15	90
03	7	8	56
04	8	9	72
05	8	7	56
06	12	10	120
07	14	15	210

Subject Identification Number	Travel Score (X)	Knowledge Score (Y)	XY
08	15	13	195
09	15	15	225
10	<u>17</u>	<u>14</u>	<u>238</u>
	$\sum X = 106$	$\sum Y = 116$	$\sum XY = 1302$
	$\sum X^2 = 1308$	$\sum Y^2 = 1434$	
	$(\sum X)^2 = 11236$	$(\sum Y)^2 = 13456$	

$$\begin{aligned}
 \text{Computation: } r &= \frac{N \sum XY - \sum X \sum Y}{\sqrt{N \sum X^2 - (\sum X)^2} \sqrt{N \sum Y^2 - (\sum Y)^2}} \\
 &= \frac{10(1302) - (106)(116)}{\sqrt{10(1308) - 11236} \sqrt{10(1434) - 13456}} \\
 &= \frac{13020 - 12296}{\sqrt{13080 - 11236} \sqrt{14340 - 13456}} \\
 &= \frac{724}{\sqrt{1844} \sqrt{884}} \\
 &= \frac{724}{1276.61} \\
 &= .567
 \end{aligned}$$

The calculation of r requires a number of arithmetic operations on the X and Y scores. $\sum X$ is simply the sum of the scores on variable X . $\sum X^2$ is the sum of the squared scores on X (each score is first squared and then the sum of the squared scores is obtained). The quantity $(\sum X)^2$ is the square of the sum of the scores: The total of the X scores ($\sum X$) is first calculated and then this total is squared. It is important not to confuse the two quantities, $\sum X^2$ and $(\sum X)^2$. The same calculations are made, using the Y scores, to obtain $\sum Y$, $\sum Y^2$, and $(\sum Y)^2$. To find $\sum XY$, each participant's X score is multiplied by the score on Y ; these values are then summed for all subjects. When these calculations have been made, r is computed using the formula for r given above.

At this point, you may wish to examine carefully the calculations shown in [Table B.2](#) to familiarize yourself with the procedures for computing r . For practice, you might try calculating r from another set of data, such as the study shown in [Table 12.1](#).

Additional Statistical Significance Tests

In [Chapter 13](#), we examined the *t* test as one example of a statistical significance test. It is appropriate when comparing two groups' responses to a continuous dependent variable. This section describes several additional statistical significance tests. Like the *t* test, these tests are used to determine the probability that the results were due to random error. All these tests use the logic of the null hypothesis discussed in [Chapter 13](#). We will first consider how to evaluate the statistical significance of Pearson *r*, and then proceed to the chi-square test and the analysis of variance (*F* test).

Significance of Correlation Coefficient *r*

To test the null hypothesis that the population correlation coefficient is 0.00, we can consult a table of critical values of *r*. [Table C.4](#) in Appendix C shows critical values of *r* for .10, .05, and .01 levels of significance. The critical value of *r* for any given study depends on the *degrees of freedom* (*df*; see [Chapter 13](#)). Degrees of freedom refers to the number of scores that are free to vary. The *df* for the significance test for *r* is $N - 2$. In our example study on travel and knowledge, the number of paired observations is 10, so the *df* = 8. For 8 degrees of freedom, the critical value of *r* at the .05 level of significance is .632 (plus or minus). The obtained *r* must be more extreme than the critical *r* to be significant.

Because our obtained *r* (from [Table B.2](#)) of .567 is closer to zero than the critical value, we do not reject the null hypothesis—even though the magnitude of *r* is fairly large. Recall the discussion of non-significant results from [Chapter 13](#). It is possible that the correlation would reach statistically significance if you used a larger sample size or more sensitive and reliable measures of the variables.

Chi-Square (χ^2)

The chi-square (Greek letter *chi*, squared) test is used when dealing with nominal scale data. It is used when the data consist of frequencies (i.e., the number of participants that fall into each of several categories). Chi-square can be used with either experimental or non-experimental data. The major requirement is that both variables are studied using nominal scales.[Page 339](#)

Example

Suppose you want to know whether there is a relationship between sex and hand dominance. To do this, you sample 50 males and 50 females and ask whether they are right-handed, left-handed, or ambidextrous (use both hands with equal skill). Your data collection involves classifying each person as male or female and as right-handed, left-handed, or ambidextrous.

Fictitious data for such a study are presented in [Table B.3](#). The frequencies labelled as “O” (for “observed”) in each of the six cells in the table refer to the number of male and female subjects who fall into each of the three hand-dominance categories. The frequencies labelled “E” (for “expected”) refer to frequencies that are expected if the null hypothesis is correct. It is important that each subject falls into only one of the cells when using chi-square (that is, no subject can be counted as both male and female or both right- and left-handed).

Table B.3 Data for Hypothetical Study on Hand Dominance: Chi-Square Test

Sex of Subject	Hand Dominance			Row Totals
	Right	Left	Ambidextrous	
Male	O ₁ = 15 E ₁ = 25	O ₂ = 30 E ₂ = 20	O ₃ = 5 E ₃ = 5	50
Female				

Sex of Subject	Hand Dominance			Row Totals
	Right	Left	Ambidextrous	
Female	O ₄ = 35 E ₄ = 25	O ₅ = 10 E ₅ = 20	O ₆ = 5 E ₆ = 5	50
Column totals	50	40	10	N = 100

Computations: Cell number $\frac{(O - E)^2}{E}$

1	4.00
2	5.00
3	0.00

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

4	4.00
5	5.00
6	<u>0.00</u>
	$\sum = 18.00$

The chi-square test examines the extent to which the frequencies that are actually observed in the study differ from the frequencies that are expected if the null hypothesis is correct. The null hypothesis states that there is no relationship between sex and hand dominance: Males and females do not differ on this characteristic.

The formula for computing chi-square is

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where O is the *observed* frequency in each cell, E is the *expected* frequency in each cell, and the symbol Σ refers to summing over all cells. The steps in calculating the value of χ^2 are as follows:

Step 1: Arrange the observed frequencies in a table such as [Table B.3](#). Note that in addition to the observed frequencies in each cell, the table presents row totals, column totals, and the total number of observations (N).

Step 2: Calculate the expected frequencies for each of the cells in the table.

The expected frequency formula is

$$E = \frac{\text{row total} \times \text{column total}}{N}$$

where the row total refers to the row total for the cell, and the column total refers to the column total for the cell. Thus, the expected frequency for cell 1 (male right-handedness) is

$$E_1 = \frac{50 \times 50}{100} = 25$$

The expected frequencies for each of the cells are shown in [Table B.3](#) below the observed frequencies.

Step 3: Calculate the quantity $(O - E)^2/E$ for each cell. For cell 1, this quantity is

$$\frac{(15 - 25)^2}{25} = \frac{100}{25} = 4.00$$

Step 4: Find the value of χ^2 by summing the $(O - E)^2/E$ values found in step 3. The calculations for obtaining χ^2 for the example data are shown in [Table B.3](#).

Significance of Chi-Square

The significance of the obtained χ^2 value can be evaluated by consulting a table of critical values of χ^2 , like [Table C.1](#) in Appendix C. The critical χ^2 values indicate the value that the *obtained* χ^2 must exceed to be significant at the .10 level, the .05 level, and the .01 level.

As is the case for all statistical tests, the critical value of χ^2 for any given study depends on the degrees of freedom. For a chi-square test, the *degrees of freedom* are the number of cells in which the frequencies are free to vary once we know the row totals and column totals. The degrees of freedom for chi-square are calculated as

$$df = (R - 1)(C - 1)$$

where R is the number of rows in the table and C is the number of columns. In our example in [Table B.3](#), there are two rows and three columns, so there are $(2 - 1)(3 - 1) = 2$ degrees of freedom. In a study with three rows and three columns, there are 4 degrees of freedom, and so on.

To use [Table C.1](#), find the correct degrees of freedom and then determine the critical value of χ^2 necessary to reject the null hypothesis at the chosen significance level. With 2 degrees of freedom, the obtained χ^2 value must be greater than the critical value of 5.991 to be significant at the .05 level. There is only a .05 probability that a χ^2 beyond 5.991 would occur if only random error is operating. Because the obtained χ^2 from our example is 18.00, we can reject the null hypothesis that there is no relationship between sex and hand dominance. (Of course, the chi-square was based on fictitious data, but you could determine for yourself whether there is a relationship by gathering and analyzing your own data.)[Page 341](#)

Concluding Remarks

The chi-square test is used frequently across the behavioural sciences. The calculational formula described is generalizable to expanded studies in which there are more categories on either of the variables. One note of caution, however: When both variables have only two categories, so that there are only two rows and two columns, the formula for calculating chi-square changes slightly. In such cases, the formula is

$$\chi^2 = \sum \frac{(|O - E| - .5)^2}{E}$$

where $|O - E|$ is the absolute value of $O - E$, and .5 is a constant that is subtracted for each cell.

Analysis of Variance (*F* Test): Overview

As noted briefly in [Chapter 13](#), the analysis of variance (ANOVA), or *F* test, is used to determine whether there is a significant difference among groups on a dependent measure that uses either an interval or ratio scale. ANOVA is a versatile test that can be adapted for various designs. Here we offer procedures for calculating *F* for the following designs: between-subjects with one or two independent variables, and within-subjects with one independent variable.

Analysis of Variance: One Independent Variable, Between-Subjects Design

To illustrate the use of the analysis of variance, let's consider a hypothetical experiment on physical distance and self-disclosure. You might predict that people will reveal more about themselves to an interviewer when they are sitting close to the interviewer than they will when sitting farther away. To test this idea, you conduct an experiment. Participants are told that interviewing techniques are being studied. Each participant is seated in a room; the interviewer comes into the room and sits at one of three distances from the participant: close (2 feet, or 0.61 metres), medium (4 feet, or 1.22 metres), or far (6 feet, or 1.83 metres). The interviewer's distance is the

independent variable manipulation. Participants are randomly assigned to one of the three distance conditions, and the interviewer's behaviour is otherwise constant in all conditions. The interview consists of a number of questions, and the dependent variable is the number of personal, revealing statements made by the participant during the interview.

Fictitious data for this between-subjects experiment are shown in [Table B.4](#). Note there are five participants in each group. The first step in calculating the F ratio is to calculate different variance estimates called *sum of squares*.

Table B.4 Data for Hypothetical Experiment on Distance and Self-Disclosure:
Analysis of Variance

Distance (A)		
Close ($A1$)	Medium ($A2$)	Far ($A3$)
33	21	20
24	25	13
31	19	15
29	27	10
<u>34</u>	<u>26</u>	<u>14</u>
$T_{A1} = 151$	$T_{A2} = 118$	$T_{A3} = 72$
$n_{A1} = 5$	$n_{A2} = 5$	$n_{A3} = 5$
$X_{A1} = 30.20$	$X_{A2} = 23.60$	$X_{A3} = 14.40$
$\sum X_{A1}^2 = 4623$	$\sum X_{A2}^2 = 2832$	$\sum X_{A3}^2 = 1090$
$T_{A1}^2 = 22801$	$T_{A2}^2 = 13924$	$T_{A3}^2 = 5184$
$SS_{TOTAL} = \sum X^2 - \frac{G^2}{N}$		
$= (4623 + 2832 + 1090) - \frac{(151 + 118 + 72)^2}{15}$		
$= 8545 - 7752.07$		
$= 792.93$		
$SS_A = \sum \frac{T_a^2}{n_a} - \frac{G^2}{N}$		
$= \left[\frac{(151)^2}{5} + \frac{(118)^2}{5} + \frac{(72)^2}{5} \right] - 7752.07$		
$= 8381.80 - 7752.07$		
$= 629.73$		
$SS_{ERROR} = \sum X^2 - \sum \frac{T_a^2}{n_a} = 8545 - 8381.80$		
$= 163.20$		

Sum of Squares

Sum of squares stands for the *sum of squared deviations from the mean*. Computing an analysis of variance for the data in [Table B.4](#) involves three sums of squares: (1) SS_{TOTAL} , the sum of squared deviations of each individual score from the grand mean; (2) SS_A , the sum of squared deviations of each of the group means from the grand mean; and (3) SS_{ERROR} , the sum of squared deviations of the individual scores from their respective group means. The “ A ” in SS_A is used to indicate that we are dealing with the systematic variance associated with independent variable A .
Page 342

All three sums of squares are deviations from a particular mean. (Recall that we calculated deviations [earlier](#) when finding the variance in a set of scores.) We could calculate the deviations directly with the data in [Table B.4](#), but

such calculations are hard to work with, so we will use simplified formulas for computational purposes. The three computational formulas are explained below. The actual computations are shown in [Table B.4](#). As you work through the example, note that $SS_{TOTAL} = SS_A + SS_{ERROR}$. The total *sum of squares* is being divided into two parts: variation attributable to the independent variable A, and variation attributable to error.

SS_{TOTAL}

The formula for SS_{TOTAL} is

$$\sum X^2 - \frac{G^2}{N}$$

where $\sum X^2$ is the sum of the squared scores of all subjects in the experiment. Each of the scores is squared first and then added. Thus, for the data in [Table B.4](#), $\sum X^2$ is $33^2 + 24^2 + 31^2$ and so on until all of the scores have been squared and added. If you are doing the calculations by hand or with a calculator, it may be convenient to find the $\sum X^2$ for the scores in each group and then add these up for your final computation. This is what we did for the data in the table. The G in the formula stands for the grand total of all of the scores. This involves adding up the scores for all participants. The grand total is then squared and divided by N , the total number of participants in the experiment. When computing the sum of squares, you should always keep the calculations clearly labelled, because you can simplify later calculations by referring to these earlier ones.[Page 343](#)

SS_A

The formula for SS_A is

$$\sum \frac{T_a^2}{n_a} - \frac{G^2}{N}$$

The T_a in this formula refers to the total of the scores in Group a of independent variable A . (T_a is a shorthand notation for $\sum X$ in each group. [Recall that we calculated $\sum X$ for the mean.] The T_a symbol is used here to avoid having to deal with too many \sum signs in our calculation procedures.) The a is used to symbolize the particular group number; thus, T_a is a general symbol for T_1 , T_2 , and T_3 . Looking at our data in [Table B.4](#), $T_1 = 151$, $T_2 = 118$, and $T_3 = 72$. These are the sums of the scores in each of the groups. After T_a has been calculated, T_a^2 is found by squaring T_a . Now, T_a^2 is divided by n_a , the number of subjects in Group a . Once the quantity T_a^2/n_a has been computed for each group, the quantities are summed as indicated by the \sum symbol.

Note that the second part of the formula, G^2/N , was calculated as a step toward SS_{TOTAL} . Because we already have this quantity, it need not be calculated again when computing SS_A . After obtaining SS_A , we can now compute SS_{ERROR} .

SS_{ERROR}

The formula for SS_{ERROR} is

$$\sum X^2 - \sum \frac{T_a^2}{n_a}$$

Both halves of this equation were calculated above in obtaining SS_{TOTAL} and SS_A . To obtain SS_{ERROR} , find these quantities and perform the proper subtraction.

To check your calculations, ensure that $SS_{TOTAL} = SS_A + SS_{ERROR}$.

Mean Squares

After obtaining the sum of squares, it is necessary to compute the *mean squares*. Mean square stands for the *mean of the sum of the squared deviations from the mean* or, more simply, the mean of the sum of squares. The mean square (MS) is the sum of squares divided by the degrees of freedom. The degrees of freedom are determined by the number of scores in the sum of squares that are free to vary. The mean squares are the variances that are used in computing the value of F . The necessary computations are shown in an analysis of variance summary table in [Table B.5](#). Constructing a summary table is the easiest way to complete the computations.

Table B.5 Analysis of Variance Summary Table

Source of Variance	Sum of Squares	df	Mean Square	F
A	SS_A	$a - 1$	SS_A/df_A	MS_A/MS_{ERROR}
Error	SS_{ERROR}	$N - a$	SS_{ERROR}/df_{ERROR}	
Total	SS_{TOTAL}	$N - 1$		
A	629.73	2	314.87	23.15
Error	163.20	12	13.60	
Total	792.93	14		

From [Table B.5](#), you can see that the mean squares that concern us are the mean square for A (systematic variance) and the mean square for error (error variance). The formulas are

$$MS_A = SS_A/df_A$$

$$MS_{ERROR} = SS_{ERROR}/df_{ERROR}$$

Page 344 where $df_A = a - 1$ (the number of groups minus one) and $df_{ERROR} = N - a$ (the total number of subjects minus the number of groups).

Obtaining the F Value

The obtained F is found by dividing MS_A by MS_{ERROR} . If only random error is operating, the expected value of F is 1.0. The greater the F value, the lower the probability that the results of the experiment were due to chance error.

Significance of F

To determine the significance of the obtained F value, it is necessary to compare the obtained F to a critical value of F . [Table C.3](#) in Appendix C shows critical values of F for significance levels of .10, .05, and .01. To find the critical value of F , locate on the table the degrees of freedom for the numerator of the ratio (the systematic variance) and the degrees of freedom for the denominator of the F ratio (the error variance). The intersection of these two degrees of freedom on the table is the critical F value.

The appropriate degrees of freedom for our sample data are 2 and 12 (see [Table B.5](#)). The critical F value from [Table C.3](#) is 3.89 for a .05 level of significance. For the results to be significant, the obtained F value must be equal to or greater than the critical value. Because the obtained value of F in [Table B.5](#) (23.15) is greater than the critical value, we conclude that the results are significant and reject the null hypothesis that the means of the groups are equal in the population.

Concluding Remarks

The analysis of variance for one independent variable with a between-subjects design can be used when there are two or more groups in the experiment. The calculations are the same whether participants are randomly assigned to condition or are grouped according to a participant variable (e.g., sex, low versus high self-esteem). The formulas are also applicable to cases in which the number of participants in each group is not equal (although you should try your best to have approximately equal numbers of participants in each group).

When the design of the experiment includes more than two levels of the independent variable (as in our example experiment, which had three groups), the obtained F value only tells us that there is a significant difference among means—but does not tell us whether any two specific groups are significantly different from one another. Follow-up tests are needed. One way to examine the difference between two groups in such a study is to adapt the formula for SS_A using only two of the groups (the df to use for the mean square would then be $2 - 1$). When calculating the F ratio in the final step, use MS_{ERROR} as the denominator. More complicated procedures for evaluating the difference between two groups in such designs are available, and easily calculated with statistical software. Page 345

Analysis of Variance: Two Independent Variables, Between-Subjects Design

In this section, we will describe the computations for analysis of variance with a factorial design containing two independent variables (see [Chapter 11](#)). The formulas apply to an $A \times B$ factorial design with any number of levels of each independent variable. The formulas apply only to a completely between-subjects design with different subjects in each group, and the number of subjects in each group must be equal. This analysis expands on the ANOVA for one independent variable we just explored, and can be adapted for use in more complicated designs (e.g., within-subjects or unequal numbers of participants across conditions). With these limitations in mind, let's consider example data from a hypothetical experiment.

The experiment uses a 2×2 IV \times PV factorial design. Variable A is the type of instruction used in a course, and variable B is students' intelligence level. The students are classified as of either "low" or "high" intelligence on the basis of intelligence test scores and are randomly assigned to one of two types of classes. One class uses the traditional lecture method; the other class uses an individualized learning approach with frequent testing over small amounts of material, tutors to help individual students, and a requirement that students master each section of material before going on to the next section. The information presented to students in the two classes is identical. At the end of the course, all students take the same test, which covers all of the material presented in the course. The score on this examination is the dependent variable.

[Table B.6](#) shows fictitious data for such an experiment, with five participants in each condition. This design allows us to evaluate three effects—the main effect of A , the main effect of B , and the $A \times B$ interaction (see [Chapter 11](#)). The main effect of A assesses whether one type of instruction is superior to the other; the main effect of B assesses whether high-intelligence students score differently on the test than do low-intelligence students; the $A \times B$ interaction examines whether the effect of one independent variable (e.g., instruction) is different depending on the particular level of the other variable (e.g., intelligence).

To compute the analysis of variance, begin by calculating the sum of squares for the following sources of variance in the data: SS_{TOTAL} , SS_A , SS_B , $SS_{A \times B}$, and SS_{ERROR} . The procedures for calculation are similar to the calculations performed for the analysis of variance with one independent variable. The numerical calculations for the example data are shown in [Table B.7](#). Next, each formula is considered in turn.

SSTOTAL

The SS_{TOTAL} is computed in the same way as the previous analysis. The formula is

$$SS_{TOTAL} = \sum X^2 - \frac{G^2}{N}$$

where $\sum X^2$ is the sum of the squared scores of all subjects in the experiment, G is the grand total of all of the scores, and N is the total number of subjects. It is usually easiest to calculate $\sum X^2$ and G in smaller steps by calculating subtotals separately for each group in the design. The subtotals are then added. This is the procedure followed in [Tables B.6](#) and [B.7](#). Page 346

Table B.6 Data for Hypothetical Experiment on the Effect of Type of Instruction and Intelligence Level on Exam Score: Analysis of Variance

	Intelligence (<i>B</i>)	
	Low (<i>B1</i>)	High (<i>B2</i>)
Traditional lecture (<i>A1</i>)	75 70 69 72 68	90 95 89 85 91
	$T_{A1B1} = 354$	$T_{A1B2} = 450$
	$\sum X_{A1B1}^2 = 25094$	$\sum X_{A1B2}^2 = 40552$
	$n_{A1B1} = 5$	$n_{A1B2} = 5$
	$\bar{X}_{A1B1} = 70.80$	$\bar{X}_{A1B2} = 90.00$
Individualized method (<i>A2</i>)	85 87 83 90 89	87 94 93 89 92
	$T_{A2B1} = 434$	$T_{A2B2} = 455$
	$\sum X_{A2B1}^2 = 37704$	$\sum X_{A2B2}^2 = 41439$
	$n_{A2B1} = 5$	$n_{A2B2} = 5$
	$\bar{X}_{A2B1} = 86.80$	$\bar{X}_{A2B2} = 91.00$
	$T_{B1} = 788$	$T_{B2} = 905$
	$n_{B1} = 10$	$n_{B2} = 10$
	$\bar{X}_{B1} = 78.80$	$\bar{X}_{B2} = 90.50$

Table B.7 Computations for Analysis of Variance with Two Independent Variables

$$\begin{aligned}
 SS_{TOTAL} &= \sum X^2 - \frac{G^2}{N} \\
 &= (25094 + 40552 + 37704 + 41439) \\
 &\quad - \frac{(354 + 450 + 434 + 455)^2}{20} \\
 &= 144789 - 143312.45 \\
 &= 1476.55
 \end{aligned}$$

$$\begin{aligned}
 SS_A &= \frac{\sum T_a^2}{n_a} - \frac{G^2}{N} \\
 &= \frac{(804)^2 + (889)^2}{10} - 143312.45 \\
 &= 143673.70 - 143312.45 \\
 &= 361.25
 \end{aligned}$$

$$\begin{aligned}
SS_B &= \frac{\sum T_b^2}{n_b} - \frac{G^2}{N} \\
&= \frac{(788)^2 + (905)^2}{10} - 143312.45 \\
&= 143996.90 - 143312.45 \\
&= 684.45
\end{aligned}$$

$$\begin{aligned}
SS_{A \times B} &= \frac{\sum T_{ab}^2}{n_{ab}} - \frac{G^2}{N} - SS_A - SS_B \\
&= \frac{(354)^2 + (450)^2 + (434)^2 + (455)^2}{5} - 143312.45 - 3 \\
&= 144639.40 - 143312.45 - 361.25 - 684.45 \\
&= 281.25
\end{aligned}$$

$$\begin{aligned}
SS_{ERROR} &= \sum X^2 - \frac{\sum T_{ab}^2}{n_{ab}} \\
&= 144789 - 144639.40 \\
&= 149.60
\end{aligned}$$

SS_A

The formula for SS_A is

$$SS_A = \frac{\sum T_a^2}{n_a} - \frac{G^2}{N}$$

where $\sum T_a^2$ is the sum of the squared totals of the scores in each of the groups of independent variable A , and n_a is the number of participants in each level of independent variable A . When calculating SS_A , the totals for each group of the A variable are obtained by considering all participants in that level of A , regardless of which condition of B the subject may be in. When we calculated SS_{TOTAL} , we already calculated G^2/N .

SS_B

The formula for SS_B is

$$SS_B = \frac{\sum T_b^2}{n_b} - \frac{G^2}{N}$$

Page 347 SS_B is calculated in the same way as SS_A . The only difference is that we are calculating group totals for independent variable B .

SS_{A × B}

The formula for $SS_{A \times B}$ is

$$SS_{A \times B} = \frac{\sum T_{ab}^2}{n_{ab}} - \frac{G^2}{N} - SS_A - SS_B$$

The sum of squares for the $A \times B$ interaction is computed by first calculating the quantity $\sum T_{ab}^2$. This involves squaring the total of the scores in each of the ab conditions (*cells*) in the experiment. In our example experiment in [Table B.6](#), there are four conditions; the interaction calculation considers all four of the groups. Each of the group

totals is squared, and added together. This sum is divided by n_{ab} , the number of subjects in each group. The other quantities in the formula for $SS_{A \times B}$ have already been calculated, so the computation of $SS_{A \times B}$ is relatively straightforward.

SS_{ERROR}

The quantities involved in the SS_{ERROR} formula have already been calculated. The formula is

$$SS_{ERROR} = \sum X^2 - \frac{\sum T_{ab}^2}{n_{ab}}$$

These quantities were calculated previously; perform the proper subtraction to complete the computation of SS_{ERROR} . Page 348

At this point, you may want to practise calculating the sums of squares using the data in [Table B.6](#). To check the calculations, make sure that $SS_{TOTAL} = SS_A + SS_B + SS_{A \times B} + SS_{ERROR}$.

After obtaining the sums of squares, the next step is to find the mean square for each of the sources of variance. The easiest way to do this is to use an analysis of variance summary table like [Table B.8](#).

Table B.8 Analysis of Variance Summary Table: Two Independent Variables

Source of Variance	Sum of Squares	df	Mean Square	F
A	SS_A	$a - 1$	SS_A/df_A	MS_A/MS_{ERROR}
B	SS_B	$b - 1$	SS_B/df_B	MS_B/MS_{ERROR}
$A \times B$	$SS_{A \times B}$	$(a - 1)(b - 1)$	$SS_{A \times B}/df_{A \times B}$	$MS_{A \times B}/MS_{ERROR}$
Error	SS_{ERROR}	$N - ab$	SS_{ERROR}/df_{ERROR}	
Total	SS_{TOTAL}			
A	361.25	1	361.25	38.64
B	684.45	1	684.45	73.20
$A \times B$	281.25	1	281.25	30.08
Error	149.60	16	9.35	
Total	1476.55	19		

Mean Square

The mean square for each of the sources of variance is the sum of squares divided by the degrees of freedom. The formulas for the degrees of freedom and the mean square are shown in the top portion of [Table B.8](#), and the computed values are shown in the bottom portion of the table.

Obtaining the F Value

The F value for each of the three sources of systematic variance (main effects for A and B , and the interaction) is obtained by dividing the appropriate mean square by the MS_{ERROR} . We now have three obtained F values and can evaluate the significance of each main effect and the interaction.

Significance of F

To determine whether an obtained F is significant, we need to find the critical value of F from [Table C.3](#) in Appendix C. For all of the F s in the analysis of variance summary table, the degrees of freedom are 1 and 16. Let's assume that a .01 significance level for rejecting the null hypothesis was chosen. The critical F at .01 for 1

and 16 degrees of freedom is 8.53. If the obtained F is larger than 8.53, we can say that the results are significant at the .01 level. By referring to the obtained F s in [Table B.8](#), you can see that the main effects and the interaction are all significant. We will leave it to you to interpret the main effect means and to graph the interaction. Review [Chapter 11](#) as needed. Page 349

Analysis of Variance: One Independent Variable, Within-subjects Design

The analysis of variance computations considered thus far have been limited to between-subjects designs. This section adapts these formulas for use in a within-subjects design with one independent variable.

Fictitious data for a hypothetical experiment using a within-subjects design are presented in [Table B.9](#). The experiment examines the effect of a job candidate's physical attractiveness on judgments of the candidate's competence. The independent variable is the candidate's physical attractiveness; the dependent variable is judged competence on a 10-point scale. Participants in the experiment view two videotapes of different women performing a mechanical aptitude task that involved piecing together a number of parts. Both women do equally well, but one is physically attractive and the other is unattractive. The order of presentation of the two tapes is counterbalanced to assess order effects. For these analyses, assume order effects have already been ruled out.

Table B.9 Data for Hypothetical Experiment on Attractiveness and Judged Competence: Within-Subjects Analysis of Variance

Subjects (or subject pairs)	Condition (A)		T_s	T_s^2
	Unattractive candidate (A_1)	Attractive candidate (A_2)		
#1	6	8	14	196
#2	5	6	11	121
#3	5	9	14	196
#4	7	6	13	169
#5	4	6	10	100
#6	3	5	8	64
#7	5	5	10	100
#8	4	7	11	121
	$T_{A1} = 39$	$T_{A2} = 52$		$\sum T_s^2 = 1067$
	$\sum X_{A1}^2 = 201$	$\sum X_{A2}^2 = 352$		
	$\bar{X}_{A1} = 4.88$	$\bar{X}_{A2} = 6.50$		

Condition (A)	
Subjects (or subject pairs)	Unattractive candidate (A_1) Attractive candidate (A_2) T_s T_s^2
$SS_{TOTAL} = \sum X^2 - \frac{G^2}{N}$	$= (201 + 352) - \frac{(39 + 52)^2}{16}$ $= 553 - 517.56$ $= 35.44$
$SS_A = \frac{\sum T_a^2}{n_a} - \frac{G^2}{N}$	$= \frac{(39)^2 + (52)^2}{8} - 517.56$ $= 528.13 - 517.56$ $= 10.57$
$SS_{SUBJECTS} = \frac{\sum T_s^2}{n_s} - \frac{G^2}{N}$	$= \frac{1067}{2} - 517.56$ $= 533.50 - 517.56$ $= 15.94$
$SS_{ERROR} = SS_{TOTAL} - SS_A - SS_{SUBJECTS}$	$= 35.44 - 10.57 - 15.94$ $= 8.93$

The main difference between the within-subjects analysis of variance and the between-subjects analysis described earlier is that the effect of participant (or *subject*) differences becomes a source of variance. There are four sources of variance in the within-subjects analysis of variance, and so four sums of squares are calculated:

$$SS_{TOTAL} = \sum X^2 - \frac{G^2}{N}$$

$$SS_A = \frac{\sum T_a^2}{n_a} - \frac{G^2}{N}$$

$$SS_{SUBJECTS} = \frac{\sum T_s^2}{n_s} - \frac{G^2}{N}$$

$$SS_{ERROR} = SS_{TOTAL} - SS_A - SS_{SUBJECTS}$$

The calculations for these sums of squares are shown in the lower portion of [Table B.9](#). Refer to earlier calculations for reminders of how to calculate most quantities in these formulas. The only new quantity involves the calculation of $SS_{SUBJECTS}$. The term refers to the squared total score of each subject—that is, the squared total of the scores that each subject gives when measured in each of the different groups in the experiment. The quantity $\sum T_s^2$ refers to the sum of these squared totals for all subjects. The calculation of $SS_{SUBJECTS}$ is completed by dividing $\sum T_s^2$ by n_s and then subtracting by G^2/N . The term n_s refers to the number of scores that each subject gives. Because our hypothetical experiment has two conditions, $n_s = 2$, the total for each subject is based on two scores.

The within-subjects analysis of variance summary table is shown in [Table B.10](#). The procedures for computing the mean squares and obtaining F are similar to our previous calculations. Note that the mean square and F for $SS_{SUBJECTS}$ are not computed. There is usually no reason to know or care whether subjects differ significantly from one another. The ability to calculate this source of variance does have the advantage of reducing the amount of error variance—in a between-subjects design, subject differences are part of the error variance. Because there is only one score per subject in the between-subjects design, it is impossible to estimate the influence of subject differences.

Table B.10 Analysis of Variance Summary Table: Within-Subjects Design

Source of Variance	Sum of Squares	df	Mean Square	F
<i>A</i>	SS_A	$a - 1$	SS_A/df_A	MS_A/MS_{ERROR}
Subjects	$SS_{SUBJECTS}$	$s - 1$	—	
Error	SS_{ERROR}	$(a - 1)(s - 1)$	SS_{ERROR}/df_{ERROR}	
Total	SS_{TOTAL}	$N - 1$		
<i>A</i>	10.57	1	10.57	8.26
Subjects	15.94	7	—	
Error	8.93	7	1.28	
Total	35.44	15		

You can use the summary table and the table of critical *F* values to determine whether the difference between the two groups is significant. The procedures are identical to those discussed previously.

Analysis of Variance: Conclusion

The analysis of variance is a very useful test that can be extended to any type of factorial design, including those that use a combination of between-subjects and within-subjects in the same design. The method of computing analysis of variance is much the same regardless of the complexity of the design. A section on analysis of variance as brief as this is not intended to cover all of the many aspects of such a general statistical technique. However, you should now have the background to compute an analysis of variance and to understand the more detailed discussions of analysis of variance in advanced statistics texts.

Effect Size

In general, measures of effect size indicate the extent to which variables are associated. These are very important measures because they help assess the strength or amount of association across studies, and can be combined in *meta-analyses* (see [Chapter 14](#)) to determine overall patterns. Different effect size measures are appropriate for different data. See Grissom and Kim (2012) for a thorough discussion of various effect size measures.

Effect Size as Strength of Association

Correlation

Recall that correlation coefficients can be interpreted as effect sizes, because they indicate the magnitude of the relationship between two variables. These numbers range from 0.00, indicating no relationship, to 1.00; correlations above .50 are considered to be indicative of very strong relationships (Cohen, 1992). In much research, expect correlations between about .15 and .40. Correlations between about .10 and .20 are considered weak, but may be statistically significant with very large sample sizes. Typically, larger correlations are more useful. Nonetheless, weak correlations might be important for theoretical or practical reasons, depending on the context.

Cramer's V to Accompany Chi-Square Tests

In addition to determining whether there is a significant relationship using the chi-square (χ^2) test, an indicator of effect size tells you the strength of association between the variables. A common effect size indicator for nominal scale data is called Cramer's V (an adaptation of the *phi* coefficient; see Grissom & Kim, 2012). The V coefficient is computed after obtaining the value of chi-square. The formula is

$$V = \sqrt{\frac{\chi^2}{N(k - 1)}}$$

In this formula, N is the total number of cases or subjects and k is the smaller of the rows or columns in the table; thus, in our [earlier](#) example with three columns (hand dominance) and two rows (sex), the value of k is 2, the lower value.

The value of V for the sex and hand dominance example in [Table B.3](#) is

$$V = \sqrt{\frac{18}{100(2 - 1)}} = \sqrt{.18} = .42$$

Values of Cramer's V are interpreted as correlation coefficients. This result of .42 offers an estimate that in the population, hand dominance and sex would be moderately related (if these data weren't fictitious!).[Page 352](#)

Effect Size as Proportion of Variance Explained

Squared Correlation and Squared Multiple Correlation

Recall from [Chapter 12](#) that we can square correlation coefficients (r) and multiple correlation coefficients (R). When squared, these values can be interpreted as a proportion of variance in one variable that is explained by the other (or by the combined set of predictors, as in the case of regression).

Omega Squared (ω^2) to Accompany ANOVA

After computing an analysis of variance and evaluating the significance of the F statistic, you need to examine effect size. One commonly used effect size measure is called *eta squared* (symbolized η^2). Although it can be easily interpreted like a correlation coefficient, this estimate is misleading unless the true population effect size is very large and the researcher has drawn a very large sample. A more robust alternative is called *omega squared* (symbolized ω^2). It can be adapted for use in between-subjects

designs with one independent variable, as well as to estimate the strength of main effects and interactions in factorial designs. The general formula is

$$\omega^2 = \frac{SS_{effect} - (df_{effect}MS_{error})}{SS_{Total} + MS_{error}}$$

In the 2×2 design described [earlier](#), the effects of teaching method (Factor A) and student intelligence (Factor B) on test scores were investigated. We could use omega squared to estimate the effect size for the main effect of Factor A. We can extract the values we need from [Table B.8](#): SS_A was 361.25, MS_{error} was 9.35, df_A was 1, and SS_{TOTAL} was 1479.55. The value of omega squared then would be

$$\begin{aligned}\omega^2 &= \frac{361.25 - (1)(9.35)}{1476.55 + 9.35} \\ &= .24\end{aligned}$$

This result would indicate that 24 percent of the variance in final exam scores was explained by differences in the two teaching methods (again, if these data weren't fictitious!).

Effect Size as a Standardized Difference between Means

Recall from [Chapter 12](#) that Cohen's d is the appropriate effect size measure when comparing two groups on a dependent variable. Cohen's d can be used to help interpret the nature of the effect, after a t test has evaluated whether two groups differ significantly on the dependent variable. This effect size measure results in a value interpreted in standard units. Calculate the value of Cohen's d using the means (M) and standard deviations (SD) of the two groups:

$$d = \frac{M_1 - M_2}{\sqrt{\frac{(SD_1^2 + SD_2^2)}{2}}}$$

Note that the formula uses M and SD instead of \bar{X} and s . These abbreviations are used in APA style (see [Appendix A](#)). Try applying this formula to the (real!) data in [Figure 13.5](#). You will see that the effect size estimating the relationship between viewing multitasking peers and lecture comprehension is very large ($d = 1.50$). In other words, sitting behind someone multitasking on a laptop during lectures decreases comprehension by an average of 1.5 standard deviations, compared to not sitting behind such a multitasker (see Sana et al., 2013).

Statistical Tables

Table C.1 Critical Values of Chi-Square

Degrees of Freedom	Probability Level		
	.10	.05	.01
1	2.706	3.841	6.635
2	4.605	5.991	9.210
3	6.251	7.815	11.345
4	7.779	9.488	13.277
5	9.236	11.070	15.086
6	10.645	12.592	16.812
7	12.017	14.067	18.475
8	13.362	15.507	20.090
9	14.684	16.919	21.666
10	15.987	18.307	23.209
11	17.275	19.675	24.725
12	18.549	21.026	26.217
13	19.812	22.362	27.688
14	21.064	23.685	29.141
15	22.307	24.996	30.578
16	23.542	26.296	32.000
17	24.769	27.587	33.409
18	25.989	28.869	34.805
19	27.204	30.144	36.191
20	28.412	31.410	37.566

Table C.2 Critical Values of t

Significance Level*				
	.05	.025	.01	.005 <i>one-tailed test</i>
<i>df</i>	.10	.05	.02	.01 <i>two-tailed test</i>
1	6.314	12.706	31.821	63.657
2	2.920	4.303	6.965	9.925
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499
8	1.860	2.306	2.896	3.355
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.169
11	1.796	2.201	2.718	3.106
12	1.782	2.179	2.681	3.055
13	1.771	2.160	2.650	3.012
14	1.761	2.145	2.624	2.977
15	1.753	2.131	2.602	2.947
16	1.746	2.120	2.583	2.921
17	1.740	2.110	2.567	2.898
18	1.734	2.101	2.552	2.878
19	1.729	2.093	2.539	2.861
20	1.725	2.086	2.528	2.845
21	1.721	2.080	2.518	2.831
22	1.717	2.074	2.508	2.819
23	1.714	2.069	2.500	2.807
24	1.711	2.064	2.492	2.797
25	1.708	2.060	2.485	2.787
26	1.706	2.056	2.479	2.779
27	1.703	2.052	2.473	2.771

Significance Level*				
	.05	.025	.01	.005 one-tailed test
df	.10	.05	.02	.01 two-tailed test
28	1.701	2.048	2.467	2.763
29	1.699	2.045	2.462	2.756
30	1.697	2.042	2.457	2.750
40	1.684	2.021	2.423	2.704
60	1.671	2.000	2.390	2.660
120	1.658	1.980	2.358	2.617
∞	1.645	1.960	2.326	2.576

* Use the top significance level when you have predicted a specific directional difference (a one-tailed test; e.g., Group 1 will be greater than Group 2). Use the bottom significance level when you have predicted only that Group 1 will differ from Group 2 without specifying the direction of the difference (a two-tailed test).

Page 355

Table C.3 Critical Values of F

Denominator (Error)	α	df for Numerator (Systematic)											
		1	2	3	4	5	6	7	8	9	10	11	12
1	.10	39.9	49.5	53.6	55.8	57.2	58.2	58.9	59.4	59.9	60.2	60.5	60.7
	.05	161	200	216	225	230	234	237	239	241	242	243	244
	.01	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056	6083	6107
2	.10	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.40	9.41
	.05	18.5	19.00	19.2	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4
	.01	98.5	99.00	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4
3	.10	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.22
	.05	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74
	.01	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	27.1
4	.10	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.91	3.90
	.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91

Denominator α (Error)	<i>df</i> for Numerator (Systematic)												
	1	2	3	4	5	6	7	8	9	10	11	12	
.01	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.4	
5	.10	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.28	3.27
	.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.71	4.68
	.01	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.96	9.89
6	.10	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.92	2.90
	.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00
	.01	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72
7	.10	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.68	2.67
	.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57
	.01	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.54	6.47
8	.10	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.52	2.50
	.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28
	.01	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.73	5.67
9	.10	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.40	2.38
	.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07
	.01	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.18	5.11
10	.10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.30	2.28
	.05	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91
	.01	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.77	4.71
11	.10	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.23	2.21
	.05	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79
	.01	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.46	4.40
12	.10	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.17	2.15
	.05	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69
	.01	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.22	4.16

Denominator <i>a</i> (Error)	<i>df</i> for Numerator (Systematic)												
	1	2	3	4	5	6	7	8	9	10	11	12	
13	.10	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.12	2.10
	.05	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60
	.01	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96
14	.10	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.08	2.05
	.05	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53
	.01	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80
15	.10	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.04	2.02
	.05	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48
	.01	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67
16	.10	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	2.01	1.99
	.05	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42
	.01	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.62	3.55
17	.10	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.98	1.96
	.05	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38
	.01	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.46
18	.10	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.96	1.93
	.05	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34
	.01	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.43	3.37
19	.10	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.94	1.91
	.05	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31
	.01	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36	3.30
20	.10	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.92	1.89
	.05	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28
	.01	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.29	3.23
22	.10	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.88	1.86
	.05	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23

Denominator α (Error)	<i>df</i> for Numerator (Systematic)											
	1	2	3	4	5	6	7	8	9	10	11	12
.01	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18	3.12
24	.10	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.85
	.05	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.21
	.01	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.09
26	.10	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.84
	.05	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18
	.01	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	3.02
28	.10	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.81
	.05	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.15
	.01	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.96
30	.10	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.79
	.05	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13
	.01	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.91
40	.10	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.73
	.05	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04
	.01	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.73
60	.10	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.68
	.05	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.95
	.01	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.56
120	.10	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.62
	.05	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.87
	.01	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.40
200	.10	2.73	2.33	2.11	1.97	1.88	1.80	1.75	1.70	1.66	1.63	1.60
	.05	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.84
	.01	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41	2.34

Denominator <i>a</i> (Error)	<i>df</i> for Numerator (Systematic)											
	1	2	3	4	5	6	7	8	9	10	11	12
∞	.10	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.57
	.05	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.79
	.01	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.25
												2.18

Page 358

Table C.4 Critical Values of *r* (Pearson Product-Moment Correlation Coefficient)

<i>df</i>	Level of Significance for Two-Tailed Test*		
	.10	.05	.01
1	.988	.997	.9999
2	.900	.950	.990
3	.805	.878	.959
4	.729	.811	.917
5	.669	.754	.874
6	.622	.707	.834
7	.582	.666	.798
8	.549	.632	.765
9	.521	.602	.735
10	.497	.576	.708
11	.476	.553	.684
12	.458	.532	.661
13	.441	.514	.641
14	.426	.497	.623
15	.412	.482	.606
16	.400	.468	.590
17	.389	.456	.575
18	.378	.444	.561
19	.369	.433	.549
20	.360	.423	.537

<i>df</i>	Level of Significance for Two-Tailed Test*		
	.10	.05	.01
25	.323	.381	.487
30	.296	.349	.449
35	.275	.325	.418
40	.257	.304	.393
45	.243	.288	.372
50	.231	.273	.354
60	.211	.250	.325
70	.195	.232	.303
80	.183	.217	.283
90	.173	.205	.267
100	.164	.195	.254

* The significance level is halved for a one-tailed test.

Writing Research Reports in APA Style

As discussed in [Chapter 2](#), *PsycINFO* is a database that allows you to find peer-reviewed journal articles in psychology and related disciplines. This appendix offers tips and strategies for using *PsycINFO* to find articles effectively.

The User Screen

The exact “look and feel” of the website you will use to search *PsycINFO* will depend on your institution’s arrangements. [Figure D.1](#) provides an example of a display; your screen will have its own appearance.

Figure D.1 A typical interface for a searchable database

The screenshot shows a search interface with a grey header bar containing links: New Search, Thesaurus, Cited References, Citation Matcher, and More. Below the header is a section titled "Choose Databases »". There is a checkbox labeled "Suggest Subject Terms". The main search area contains two rows of search fields. The first row has a text input field containing "procrastination", a dropdown menu set to "in KW Keywords", and a "Search" button. The second row has a dropdown menu set to "AND" followed by an empty text input field, a dropdown menu set to "in Select a Field (optional)", and an "Add Row" link. At the bottom of the interface are links for Basic Search, Advanced Search, Visual Search, and Search History.



TRY IT OUT!

The best way to learn how to conduct a search is to actually conduct a search! Go to your institution’s library website and find the *PsycINFO* database, keeping in mind that you might need to log in or use a virtual private network if you are off campus. Follow along with the search below to learn how to use *PsycINFO* to find psychological research.

Specifying Search Terms and Finding Results

Your most important task is to specify the search terms you want *PsycINFO* to use. These are typed into an input box. In most simple searches, such as the one shown in [Figure D.1](#), you have some other options. For example, you can limit your search to articles that have a specific word or phrase in the title.

How do you know what terms to use? Most commonly, you will want to use standard psychological terms. Identifying the term that will yield results you are seeking may take many tries. To help you, consult the *Thesaurus of Psychological Index Terms*. This thesaurus lists all the standard terms that are used to index the abstracts, and it can be accessed directly within most *PsycINFO* systems. Also, your institution's subject librarian might be able to help you generate the best terms.[Page 360](#)

Suppose you are interested in the topic of procrastination. Entering *procrastination* into the *Thesaurus of Psychological Index Terms* reveals that *motivation* is a broader term that could be relevant to your search too. If you click on *motivation*, you will see many related terms, including *achievement motivation*, *temptation*, and *procrastination*. Let's assume that you are using a standard search window as in [Figure D.1](#). Give the command to start the search for *procrastination*, and the results will be displayed.

Below, we have included some of the output of one of the articles found with a search on *procrastination*. The exact appearance of your output will depend on the computer system you are using as well as the information that you choose to display. The default output includes citation information along with the abstract itself. Some of the extra information shows how information is organized in the database (e.g., the numbers in the “population” section are codes). When you do the search, some fields will appear as hyperlinks to lead you to other information in your library

database or to other websites. Systems are continually being upgraded to enable users to more easily obtain full-text access to the articles and find other articles on similar topics.

Notice that the output is organized into “fields” of information. The full name of each field is included here; many systems allow abbreviations. You will almost always want to see the *Title* (abbreviated as TI), *Author* (AU), *Source* (SO), and *Abstract* (AB). Note additional fields such as *Publication Type* (PT), *Keywords* (KW) to briefly describe the article, and *Age Group* of participants. The *Digital Object Identifier* (DOI) field can be helpful for finding full-text sources of the article, and is included in the latest APA style referencing format mentioned in [Chapter 2](#) and [Appendix A](#).

Title:	I forgive myself, now I can study: How self-forgiveness for procrastinating can reduce future procrastination.
Author(s):	Wohl, Michael J. A., Carleton University, Department of Psychology, Ottawa, ON, Canada Pychyl, Timothy A., Carleton University, Department of Psychology, Ottawa, ON, Canada Bennett, Shannon H., Carleton University, Department of Psychology, Ottawa, ON, Canada
Address:	Wohl, Michael J. A., Carleton University, Department of Psychology, 1125 Colonel By Drive, B550 Loeb Building, Ottawa, ON, Canada, K1S 5B6
Source:	Personality and Individual Differences, Vol 48(7), May, 2010. pp. 803–808.
Publisher:	Netherlands: Elsevier Science.
Digital Object Identifier:	10.1016/j.paid.2010.01.029
Language:	English
Keywords:	self-forgiveness; procrastination; university students; psychology education
Abstract:	In the present study, we examined the association

between forgiving the self for a specific instance of procrastination and procrastination on that same task in the future. A sample of 119 first-year University students (49 male, 70 female) completed measures of procrastination and self-forgiveness immediately before each of two midterm examinations in their introductory psychology course. Results revealed that among students who reported high levels of self-forgiveness for procrastinating on studying for the first examination, procrastination on preparing for the subsequent examination was reduced. This relationship was mediated by negative affect, such that increased self-forgiveness reduced procrastination by decreasing negative affect. Results are discussed in relation to the impact of procrastination on self-directed negative affect. (PsycINFO Database Record (c) 2010 APA, all rights reserved)

Subjects: *Forgiveness; *Procrastination; College Students; Psychology Education

Classification: Personality Traits & Processes (3120)

Population: Human (10)
Male (30)
Female (40)

Location: Canada

Age Group: Adulthood (18 yrs & older) (300)

Methodology: Empirical Study; Quantitative Study

Publication Type: Journal, Peer Reviewed Journal

Number of Citations in: 32

Source:

Database: PsycINFO

Combining Search Terms and Narrowing Results

When you do a simple search with a single word or a phrase such as *procrastination*, the default search yields articles that have that word or phrase anywhere in any of the fields listed. This strategy can produce too many articles, including articles that are not directly relevant to your interests. One way to narrow the search is to limit it to certain fields. The simple search screen (see [Figure D.1](#)) may allow you to limit the search to one field, such as the title of the article. For example, you could specify *procrastination in TI* to limit your search to articles that have the term in the title of the article. Also note that there is often a “Set Limits” option within *PsycINFO*. This allows you to easily specify that the search should find only peer-reviewed journal articles (not books or doctoral dissertations) or include participants from certain age groups.

Most *PsycINFO* systems have advanced search functions that enable you to use the Boolean operators AND and OR and NOT. These can be typed as discussed below, but the advanced search screen uses prompts to help you design the search. Suppose you want to restrict the *procrastination in KEYWORD* search to students only. You can do this by asking for (*procrastination in KW*) AND (*students*), all in the same search box, or in different search boxes if they appear. The AND forces both conditions to be true for an article to be included. The parentheses are used to separate different parts of your search specification and are useful when your searches become increasingly complicated. They could have been left out of this search but are included here for illustration.

The OR operation is used to expand a search that is too narrow. Suppose you want to find journal articles that discuss romantic relationships on the Internet. The last time we checked, a quick search for *Internet AND romance* resulted in 146 citations; when you select “limit to scholarly (peer

reviewed) journals” that number dropped to 86 peer-reviewed articles. Changing the specification to *Internet AND (romance OR dating OR love)* yielded 346 peer-reviewed journal articles. Articles that have the term *Internet* as well as *any* of the other terms specified were included in the second search. Page 362

The NOT operation will exclude abstracts based on a criterion you specify, and is useful when you anticipate that the search criteria will be met by some irrelevant abstracts. In the Internet example, it is possible that the search will include articles on child predators. To exclude the term *child* from the search results, the following adjustment can be made: *Internet AND (romance OR dating OR love) NOT child*. This search resulted in 318 abstracts instead of the 346 obtained previously.

Another helpful search tool is the wildcard asterisk (*). The asterisk stands for any set of letters in a word and so it can expand your search. Consider the word *romance* in the search above—by using *roman**, the search will expand to include both *romance* and *romantic*. The wildcard can be very useful with the term *child** to find *child*, *children*, *childhood*, and so on. You have to be careful when doing this, however; the *roman** search would also find *Romania* and *romanticism*. In this case, it might be more efficient to simply add *OR romantic* to the search. These search strategies are summarized in [Table D.1](#).

Table D.1 Some *Psycinfo* Search Strategies

Strategy 1: Use fields such as TI and AU.

Example: [(divorce) in TI] requires that a term appear in the title.

Strategy 2: Use AND to limit search.

Example: [divorce AND child] requires both terms to be included.

Strategy 3: Use OR to expand search.

Example: [divorce OR breakup] includes both terms.

Strategy 4: Use NOT to exclude search terms.

Example: [shyness NOT therapy] excludes any shyness articles that have the term therapy.

Strategy 5: Use the wildcard asterisk (*).

Example: [child*] finds any word that begins with child (children, childhood, etc.).

Give careful thought to your search terms. Consider the case of a student who decided to do a paper on the topic of “road rage.” She wanted to know what might cause drivers to become so angry at other drivers that they will become physically aggressive. A search on the term *road rage* led to a number of interesting articles. However, when looking at the results from the search she noticed that the major keywords (KW) included *driving behaviour* and *anger* but not *road rage*. When she asked about this, we realized that she had found only articles that included the term *road rage* in the title or abstract. This term has become popular but it may not be used in all scientific studies of the topic. She then expanded the search to include *driving AND anger* and also *dangerous driving*. The new search yielded many articles not found in the original search.

Saving Results

When you complete your search, you can e-mail the results, save them to a flash drive, or upload them to a reference manager (e.g., Zotero, RefWorks, or EndNote). When you save, you can often choose which of the fields to display. Some systems and reference managers allow you to print, copy, or save results in different formats, including APA style. Carefully note all the options available to you. When relying on automatically generated APA style references, be sure to double-check them.

Finding just the right articles can take time. Practise using the strategies in this appendix as well as other tips in [Chapter 2](#), and over time your searches will become more efficient.

Constructing a Latin Square

Use a Latin square to determine order controls for most order effects without having to include all possible orders. A Latin square to determine the orders of any N number of conditions will have N arrangements of orders in a within-subjects design (see [Chapter 8](#)). Thus, if there are four conditions, there will be four orders in a 4×4 Latin square; eight conditions will produce an 8×8 Latin square. The method for constructing a Latin square shown below will produce orders in which (1) each condition or group appears once at each order and (2) each condition precedes and follows every other condition one time.

Imagine you have been hired by a new company to design its website. You design a within-subjects study to investigate the effect of different font types on ease of reading on a computer screen. Your four conditions are as follows: Times New Roman, Calibri, Arial, and Lucida Handwriting. A Latin square for these four conditions is shown in [Figure E.1](#). Each row in the square is one of the orders of the conditions (the conditions are labelled A, B, C, and D).

Figure E.1 A Latin square with four conditions

	Order of Conditions			
	1	2	3	4
Row 1	A (Times New Roman)	B (Calibri)	D (Lucida Handwriting)	C (Arial)
Row 2	B (Calibri)	C (Arial)	A (Times New Roman)	D (Lucida Handwriting)
Row 3	C (Arial)	D (Lucida Handwriting)	B (Calibri)	A (Times New Roman)

Order of Conditions

	1	2	3	4
Row 4	D (Lucida Handwriting)	A (Times New Roman)	C (Arial)	B (Calibri)
Row 4				

Note: The four conditions were randomly given the letter designations. A = Times New Roman, B = Calibri, C = Arial, and D = Lucida Handwriting. Each row represents a different order of running the conditions.

Page 364 Use the following procedures for generating a Latin square when there is an even number of conditions:

1. Determine the number of conditions. Use letters of the alphabet to represent your N conditions: ABCD for four conditions, ABCDEF for six conditions, and so on.
2. Determine the order for the first row, using the following ordering:

$$A, B, L, C, L - 1, D, L - 2, E$$

and so on. L stands for the last or final treatment. Thus, if you have four conditions (ABCD), your order will be

$$A, B, D, C$$

With six conditions (ABCDEF), the order will be

$$A, B, F, C, E, D$$

because F is the final treatment (L), and E is the next to final treatment ($L - 1$).

3. Determine the order for the second row by increasing one letter at each position of the first row. The last letter cannot be increased, of course, so it reverts to the first letter. With six conditions, the order of the second row becomes

$$B, C, A, D, F, E$$

4. Continue this procedure for the third and subsequent rows. For the third row, increase one letter at each position of the second row:

C, D, B, E, A, F

The final 6×6 Latin square will be

A	B	F	C	E	D
B	C	A	D	F	E
C	D	B	E	A	F
D	E	C	F	B	A
E	F	D	A	C	B
F	A	E	B	D	C

5. Randomly assign each of your conditions to one of the letters to determine which condition will be in the A position, the B position, and so on.

If you have an odd number of conditions, you must make two Latin squares. For the first square, simply follow the procedures we have shown. Now create a second square that reverses the first one; that is, in each row, the first condition becomes the last, the second condition is next to last, and so on. Join the two squares together to create the final Latin square (actually a rectangle!). Thus, if there are five conditions, you will have ten possible orders to run in your study.

When you conduct your study using the Latin square to determine order, you need at least one participant per row. Usually, you will have two or more participants per row; the number of participants run in each order must be equal.

References

1. Akins, C. K., Panicker, S., & Cunningham, C. L. (2004). *Laboratory animals in research and teaching: Ethics, care, and methods*. Washington, DC: American Psychological Association.
2. Aknin, L. B., Barrington-Leigh, C. P., Dunn, E. W., Helliwell, J. F., Burns, J., Biswas-Diener, R., . . . Norton, M. I. (2013). Prosocial spending and well-being: Cross-cultural evidence for a psychological universal. *Journal of Personality and Social Psychology*, 104, 635–652.
3. Aknin, L. B., Broesch, T., Hamlin, J. K., & Van de Vondervoort, J. W. (2015). Prosocial behavior leads to happiness in a small-scale rural society. *Journal of Experimental Psychology: General*, 144, 788–795
4. Aknin, L. B., Hamlin, J. K., & Dunn, E. W. (2012). Giving leads to happiness in young children. *PLoS ONE*, 7, e39211. doi:10.1371/journal.pone.0039211
5. Aknin, L. B., Norton, M. I., & Dunn, E. W. (2009). From wealth to well-being? Money matters, but less than people think. *The Journal of Positive Psychology*, 4, 523–527.
6. Albergotti, R., & Dwoskin, E. (2014, 30 June). Facebook study sparks soul-searching and ethical questions. *Wall Street Journal* (online). Retrieved from <http://search.proquest.com.ezproxy.library.ubc.ca/docview/1541549641?accountid=14656>
7. Albright, L., & Malloy, T. E. (2000). Experimental validity: Brunswik, Campbell, Cronbach and enduring issues. *Review of General Psychology*, 4, 337–353.

8. Allen, J. P., Schad, M. M., Oudekerk, B., & Chango, J. (2014). What ever happened to the “cool” kids? Long-term sequelae of early adolescent pseudomature behavior. *Child Development*, 85, 1866–1880.
9. Alogna, V. K., Attaya, M. K., Aucoin, P., Bahnik, S., Birch, S., Birt, A. R., . . . Zwaan, R. A. (2014). Registered replication report: Schooler & Engstler-Schooler (1990). *Perspectives on Psychological Science*, 9, 556–578.
10. American Psychological Association. (2010a). *Ethical principles of psychologists and code of conduct*. Retrieved from <http://www.apa.org/ethics/code>
11. American Psychological Association. (2010b). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.
12. American Psychological Association. (2012a). *APA style guide to electronic resources* (6th ed.). Washington, DC: American Psychological Association.
13. American Psychological Association. (2012b). *Guidelines for ethical conduct in the care and use of animals*. Retrieved from <http://www.apa.org/science/leadership/care/guidelines.aspx>
14. Anderson, C. A., & Anderson, D. C. (1984). Ambient temperature and violent crime: Test of the linear and curvilinear hypotheses. *Journal of Personality and Social Psychology*, 46, 91–97.
15. 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.
16. Anderson, M. S., Ronning, E. A., DeVries, R., & Martinson, B. C. (2010). Extending the Mertonian norms: Scientists’ subscription to norms of research. *Journal of Higher Education*, 81, 366–393.

17. Andrade, B. F., & Tannock, R. (2013). The direct effects of inattention and hyperactivity/impulsivity on peer problems and mediating roles of prosocial and conduct problem behaviors in a community sample of children. *Journal of Attention Disorders*, 17, 670–680.
18. Arim, R. G., Dahinten, V. S., Marshall, S. K., & Shapka, J. D. (2011). An examination of the reciprocal relationships between adolescents' aggressive behaviors and their perceptions of parental nurturance. *Journal of Youth and Adolescence*, 40, 207–220.
19. Aronson, E., Brewer, M., & Carlsmith, J. M. (1985). Experimentation in social psychology. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (3rd ed.). New York: Random House.
20. Asbridge, M., Mann, R. E., Smart, R. G., Stoduto, G., Beirness, D., Lamble, R., & Vingilis, E. (2009). The effects of Ontario's administrative driver's licence suspension law on total driver fatalities: A multiple time series analysis. *Drugs: Education, Prevention, Policy*, 16, 140–151.
21. Assaad, J.-M., Pihl, R. O., Séguin, J. R., Nagin, D. S., Vitaro, F., & Tremblay, R. E. (2006). Intoxicated behavioral disinhibition and the heart rate response to alcohol. *Experimental and Clinical Psychopharmacology*, 14, 377–388.
22. Baddeley, A. (2003). Working memory: looking back and looking forward. *Nature Reviews Neuroscience*, 4, 829–839.
23. Bakeman, R., & Gottman, J. M. (1986). *Observing interaction*. Cambridge, UK: Cambridge University Press.
24. Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, 43, 666–678.
25. Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, 7, 543–554.

26. Bamberger, M., Rugh, J., Church, M., & Fort, L. (2004). Shoestring evaluation: Designing impact evaluations under budget, time and data constraints. *American Journal of Evaluation*, 25, 5–37.
27. Banaji, M. (2010, December). Wikipedia is the encyclopedia that anybody can edit. But have you? *Observer*, 23. Retrieved from www.psychologicalscience.org/index.php/publications/observerPage RE-2
28. Banaji, M. (2011, February). Harnessing the power of Wikipedia for scientific psychology: A call to action. *Observer*, 24. Retrieved from www.psychologicalscience.org/index.php/publications/observer
29. Bandstra, N. F., Chambers, C. T., McGrath, P. J., & Moore, C. (2011). The behavioural expression of empathy to others' pain versus others' sadness in young children. *Pain*, 152, 1074–1082.
30. Banerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics*, 27, 3–23.
31. Bargh, J. A., Chen, M. A., & Burrows, L. (1996). Automaticity of social behaviour: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, 71, 230–244.
32. Barha, C. K., Pawluski, J. L., & Galea, L. A. M. (2007). Maternal care affects male and female offspring working memory and stress reactivity. *Physiology & Behavior*, 92, 939–950.
33. Barlow, D. H., Nock, M. K., & Hersen, M. (2009). *Single case experimental designs: Strategies for studying behavior change* (3rd ed). Boston: Allyn & Bacon.
34. Barnoy, S., Ofra, L., & Bar-Tal, Y. (2012). What makes patients perceive their health care worker as an epistemic authority? *Nursing Inquiry*, 19, 128–133.

35. 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.
36. Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The “Reading the Mind in the Eyes” test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry*, 42, 241–251.
37. Bartneck, C., Duenser, A., Moltchanova, E., & Zawieska, K. (2015). Comparing the similarity of responses received from studies in Amazon’s Mechanical Turk to studies conducted online and with direct recruitment. *PLoS ONE*, 10, e0121595.
38. Baruss, I., & Rabier, V. (2014). Failure to replicate retrocausal recall. *Psychology of Consciousness: Theory, Research, and Practice*, 1, 82–91.
39. Baum, A., Gachtel, R. J., & Schaeffer, M. A. (1983). Emotional, behavioral, and psychological effects of chronic stress at Three Mile Island. *Journal of Consulting and Clinical Psychology*, 51, 565–572.
40. Baumrind, D. (1964). Some thoughts on ethics of research: After reading Milgram’s “behavioral study of obedience.” *American Psychologist*, 19, 421–423.
41. Beach, F. A. (1950). The snark was a boojum. *American Psychologist*, 5, 115–124.
42. Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavioral Research*, 43, 800–813.
43. Bem, D. J. (1981). Writing the research report. In L. H. Kidder (Ed.), *Research methods in social relations*. New York: Holt, Rinehart & Winston.

44. Bem, D. J. (2003). *Writing the empirical journal article*. Retrieved from <http://dbem.ws/WritingArticle2.pdf>
45. Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100, 407–425.
46. Bender, J. L., Jimenez-Marroquin, M.-C., & Jadad, A. R. (2011). Seeking support on Facebook: A content analysis of breast cancer groups. *Journal of Medical Internet Research*, 13, 221–232.
47. Berry, J. W. (2013). Achieving a global psychology. *Canadian Psychology*, 54, 55–61.
48. Bhattacharjee, Y. (2013, 28 June). Stapel gets community service for fabricating studies. *Science News*. Retrieved from <http://news.sciencemag.org/europe/2013/06/stapel-gets-community-service-fabricating-studies>
49. Blacklock, K., & Perry, A. (2010). Testing the application of benchmarks for children in Ontario's IBI program: Six case studies. *Journal on Developmental Disabilities*, 16, 33–43.
50. Blatchley, B., & O'Brien, K. R. (2007). Deceiving the participant: Are we creating the reputational spillover effect? *North American Journal of Psychology*, 9, 519–534.
51. Bonnet, D. G. (2012). Replication-extension studies. *Current Directions in Psychological Science*, 21, 409–412.
52. Boucher, C. M., & Scoboria, A. (2015). Reappraising past and future transitional events: The effects of mental focus on present perceptions of personal impact and self-relevance. *Journal of Personality*, 83, 361–375.
53. Boucher, H., & Ryan, C. A. (2011). Performance stress and the very young musician. *Journal of Research in Music Education*, 58, 329–345.

54. Boucher, M. E., Lecours, S., Philippe, F. L., & Arseneault, S. (2013). Parental socialization of emotion and depression in adulthood: The role of attitudes toward sadness. *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology*, 63, 15–23.
55. Bowker, A., Boekhoven, B., Nolan, A., Bauhaus, S., Glover, P., Powell, T., & Taylor, S. (2009). Naturalistic observations of spectator behavior at youth hockey games. *Sport Psychologist*, 23, 301–316.
56. Bowman, L. L., Levine, L. E., Waite, B. M., & Gendron, M. (2010). Can students really multitask? An experimental study of instant messaging while reading. *Computers & Education*, 54, 927–931.
57. Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*. Thousand Oaks, CA: Sage.
58. Brandt, M. J., Ijzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., . . . van't Veer, A. (2014). The Replication Recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, 50, 217–224.
59. Brankley, A. E., & Rule, N. O. (2014). Threat perception: How psychopathy and Machiavellianism relate to social perceptions during competition. *Personality and Individual Differences*, 71, 103–107.
60. Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, 9, 333–342.
61. Broberg, A. G., Wessels, H., Lamb, M. E., & Hwang, C. P. (1997). Effects of day care on the development of cognitive abilities in 8-year-olds: A longitudinal study. *Developmental Psychology*, 33, 62–69.
62. Brogden, W. J. (1962). The experimenter as a factor in animal conditioning. *Psychological Reports*, 11, 239–242. Page RE-3
63. Brooks, A. M., Ottley, K. M., Arbuthnott, K. D., & Sevigny, P. (2017). Nature-related mood effects: Season and type of nature contact.

Journal of Environmental Psychology, 54, 91–102.

64. Brown, M. (2008). Student perceptions of teaching evaluations. *Journal of Instructional Psychology*, 35, 177–181.
65. Bruchmüller, K., Margraf, J., & Schneider, S. (2012). Is ADHD diagnosed in accord with diagnostic criteria? Overdiagnosis and influence of client gender on diagnosis. *Journal of Consulting and Clinical Psychology*, 80, 128–238.
66. Buehler, R., Griffin, D., & Ross, M. (2002). Inside the planning fallacy: The causes and consequences of optimistic time predictions. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 250–270). Cambridge: Cambridge University Press.
67. Buehler, R., Peetz, J., & Griffin, D. (2010). Finishing on time: When do predictions influence completion times? *Organizational Behavior and Human Decision Processes*, 111, 23–32.
68. Buffardi, E. L. & Campbell, W. K. (2008). Narcissism and social networking web sites. *Personality & Social Psychology Bulletin*, 34, 1303–1314.
69. Buoy, T., & Nicoladis, E. (2018). The considerateness of codeswitching: A comparison of two groups of Canadian French-English bilinguals. *Journal of Intercultural Communication Research*, 47, 361–373.
70. Burger, J. (2009). Replicating Milgram: Would people still obey today? *American Psychologist*, 64, 1–11.
71. Bushman, B. J., & Wells, G. L. (2001). Narrative impressions of the literature: The availability bias and the corrective properties of meta-analytic approaches. *Personality and Social Psychology Bulletin*, 27, 1123–1130.

72. Buunk, A. P., Park, J. H., & Duncan, L. A. (2010). Cultural variation in parental influence on mate choice. *Cross-Cultural Research*, 44, 23–40.
73. Byrne, B. M., & van de Vijver, F. J. R. (2010). Testing for measurement and structural equivalent in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing*, 10, 107–132.
74. Cacioppo, J. T., Tassinary, L. G., & Berntson, G. G. (Eds.). (2007). *Handbook of psychophysiology* (3rd ed.). New York: Cambridge University Press.
75. Cain, P. (2013, October 2). National Household Survey: More than 90% of Kelowna neighbourhoods would have been excluded from results if Statistics Canada hadn't dropped its standards. *Global News*. Retrieved from <http://globalnews.ca/news/873012/who-filled-out-the-national-household-survey-and-why-did-statscan-cut-its-census-standards-in-half/#statscan>
76. Cain, P., & Mehler Paperny, A. (2013, October 21). Fraser Health may cut programs to pay for data Statscan didn't get. *Global News*. Retrieved from <http://globalnews.ca/news/911490/statscans-national-household-survey-missed-people-on-social-assistance-heres-why-thats-a-problem/>
77. Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., . . . & Altmejd, A. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2, 637–644.
78. Cameron, J. J., Holmes, J. G., & Vorauer, J. D. (2009). When self-disclosure goes awry: Negative consequences of revealing personal failures for lower self-esteem individuals. *Journal of Experimental Social Psychology*, 45, 217–222.
79. Cameron, J. J., Stinson, D. A., Gaetz, R., & Balchen, S. (2010). Acceptance is in the eye of the beholder: Self-esteem and motivated

- perceptions of acceptance from the opposite sex. *Journal of Personality and Social Psychology*, 99, 513–529.
80. Campbell, D. T. (1969). Reforms as experiments. *American Psychologist*, 24, 409–429.
81. Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.
82. Canadian Council on Animal Care. (2006). *Terms of Reference for Animal Care Committees*. Retrieved from http://www.ccac.ca/Documents/Standards/Policies/Terms_of_reference_for_ACC.pdf
83. Canadian Council on Animal Care. (2013). *Annual Report 2012–2013 Canadian Council on Animal Care*. Retrieved from http://www.ccac.ca/en/publications/annual_reports
84. Canadian Council on Animal Care. (2014). *About the Canadian Council on Animal Care*. Retrieved from <http://www.ccac.ca/en/about>
85. Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, and Social Sciences and Humanities Research Council of Canada. (2010). *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans*. Retrieved from <http://www.pre.ethics.gc.ca>
86. Canadian Psychological Association (2000). *Canadian Code of Ethics for Psychologists* (3rd ed.). Canadian Psychological Association: Ottawa, Canada. Retrieved from <http://www.cpa.ca/aboutcpa/committees/ethics/codeofethics/>
87. Carroll, M. E., & Overmier, J. B. (Eds.). (2001). *Animal research and human health: Advancing human welfare through behavioral science*. Washington, DC: American Psychological Association.

88. 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 behavioural testing. *Computers in Human Behavior*, 29, 2156–2160.
89. Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science*, 9, 40–48.
90. Chambers, T. (2006). What I hear you saying is . . . : Analysis of student comments from the NSSE. *College Student Journal*, 44, 3–24.
91. Champely, S. (2018). pwr: Basic Functions for Power Analysis. R package version 1.2-2. <https://CRAN.R-project.org/package=pwr>
92. Chan, M. E., & Arvey, R. D. (2012). Meta-analysis and the development of knowledge. *Perspectives on Psychological Science*, 7, 79–92.
93. Chandler, J., & Schwarz, N. (2009). How extending your middle finger affects your perception of other: Learned movements influence concept accessibility. *Journal of Experimental Social Psychology*, 45, 123–128.
94. Chapman, H. A., Bernier, D., & Rusak, B. (2010). MRI-related anxiety levels change within and between repeated scanning sessions. *Psychiatry Research: Neuroimaging*, 182, 160–164. Page RE-4
95. Chartrand, T. L., & Bargh, J. A. (2000). The mind in the middle: A practical guide to priming and automaticity research. In Reis, H. T., & Judd, C. M. (Eds.), *Handbook of research methods in social and personality psychology* (pp. 253–285). Cambridge, UK: Cambridge University Press.
96. Chopra, K. K., Ravindran, A., Kennedy, S. H., Mackenzie, B., Matthews, S., Anisman, H., . . . Levitan, R. D. (2009). Sex differences in hormonal responses to a social stressor in chronic major depression. *Psychoneuroendocrinology*, 34, 1235–1241.

97. Christensen, L. (1988). Deception in psychological research: When is its use justified? *Personality and Social Psychology Bulletin*, 14, 664–675.
98. Chung-Fat-Yim, A., Cilento, E., Piotrowska, E., & Mar, R. A. (2019). Are stories just as transporting when not in your native tongue? *Language & Cognition*, 11, 282–309.
99. Cialdini, R. B. (2008). *Influence: Science and practice* (5th ed.). Boston: Allyn & Bacon.
100. Cianci, A. M., Klein, H. J., & Seijts, G. H. (2010). The effect of negative feedback on tension and subsequent performance: The main and interactive effects of goal content and conscientiousness. *Journal of Applied Psychology*, 95, 618–630.
101. Cohen, D., Nisbett, R. E., Bowdle, B. F., & Schwarz, N. (1996). Insult, aggression, and the southern culture of honor: An “experimental ethnography.” *Journal of Personality and Social Psychology*, 70, 945–960.
102. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
103. Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155–159.
104. Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49, 997–1003.
105. Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton-Mifflin.
106. Cory, P. (2001). *The inquiry regarding Thomas Sophonow: The investigation, prosecution and consideration of entitlement to compensation*. Winnipeg: Manitoba Justice.

107. Costa, M. (2010). Interpersonal distances in group walking. *Journal of Nonverbal Behavior*, 34, 15–26.
108. Costa, P. T., Jr., & McCrae, R. R. (1985). *The NEO Personality Inventory manual*. Odessa, FL: Psychological Assessment Resources.
109. Crawford, F. (2000). Researcher in consumer behavior looks at attitudes of gratitude that affect gratuities. *Cornell Chronicle*. Retrieved from <http://www.news.cornell.edu/Chronicle/00/8.17.00/Lynn-tipping.html>
110. Creswell, J. W. (2007). *Qualitative inquiry and research design: Choosing among five approaches* (2nd ed.). Thousand Oaks, CA: Sage.
111. Creswell, J. W., & Miller, D. L. (2000). Determining validity in qualitative inquiry. *Theory into Practice*, 39, 124–130.
112. Creswell, J. W., Hanson, W. E., Clark, V. L. P., & Morales, A. (2007). Qualitative research designs: Selection and implementation. *The Counseling Psychologist*, 35, 236–264.
113. Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE*, 8, e57410.
114. Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, 3, 286–300.
115. Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.
116. Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25, 7–29.
117. Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist*, 60,

170–180.

118. Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, 66, 4–19.
119. Cutting, J. E., & Candan, A. (2015). Shot durations, shot classes, and the increased pace of popular movies. *Projections*, 9, 40–62.
120. Cuttler, C., & Graf, P. (2007). Personality predicts prospective memory task performance: An adult lifespan study. *Scandinavian Journal of Psychology*, 48, 215–231.
121. Czaja, J., Hartmann, A. S., Rief, W., & Hilbert, A. (2011). Mealtime family interactions in home environments of children with loss of control eating. *Appetite*, 56, 587–593.
122. Darredeau, C., & Barrett, S. P. (2010). The role of nicotine content information in smokers' subjective responses to nicotine and placebo inhalers. *Human Psychopharmacology*, 25, 577–581.
123. Davidson, E. J. (2005). *Evaluation Methodology Basics: The Nuts and Bolts of Sound Evaluation*. Thousand Oaks, CA: Sage Publications.
124. Denmark, F., Russo, N. P., Frieze, I. H., & Sechzer, J. A. (1988). Guidelines for avoiding sexism in psychological research: A report of the Ad Hoc Committee on Nonsexist Research. *American Psychologist*, 43, 582–585.
125. Dickstein, S., Hayden, L. C., Schiller, M., Seifer, R., & San Antonio, W. (1994). *Providence family study mealtime family interaction coding system*. Unpublished classification manual, Bradley Research Center, East Providence, RI.
126. Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, 6, 274–290.

127. Dillman, D. A. (2000). *Mail and Internet surveys: The tailored design method* (2nd ed.). New York: Wiley.
128. Dolhanty, J., & Greenberg, L. S. (2009). Emotion-focused therapy in a case of anorexia nervosa. *Clinical Psychology and Psychotherapy*, 16, 366–382.
129. Doliński, D., Grzyb, T., Folwarczny, M., Grzybała, P., Krzyszycza, K., Martynowska, K., & Trojanowski, J. (2017). Would you deliver an electric shock in 2015? Obedience in the experimental paradigm developed by Stanley Milgram in the 50 years following the original studies. *Social Psychological and Personality Science*, 8, 927–933.
130. Donner, D. D., Snowden, D. A., & Friesen, W. V. (2001). Positive emotions in early life and longevity: Findings from the Nun Study. *Journal of Personality and Social Psychology*, 80, 804–813.
131. Doyen, S., Klein, O., Pichon, C.-L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE*, 7, e29081. Page RE-5
132. Drankiewicz, D., & Dundes, L. (2003). Handwashing among female college students. *American Journal of Infection Control*, 31, 67–71.
133. Drews, F., Pasupathi, M., & Strayer, D. (2008). Passenger and cell phone conversations in simulated driving. *Journal of Experimental Psychology: Applied*, 14, 392–400.
134. Dubois-Comtois, K., & Moss, E. (2008). Beyond the dyad: Do family interactions influence children's attachment representations in middle childhood? *Attachment & Human Development*, 10, 415–431.
135. Dubue, J. D., Cheng, J. C., Vuong, W., & Westbury, C. (2018). Peer Edmonton Empathy Recruitment Scale (PEERS): A tool for student peer support worker selection and empathy measurement. *Canadian Journal of Counselling and Psychotherapy*, 52, 180–193.

136. Duggan, K. A., Reynolds, C. A., Kern, M. L., & Friedman, H. S. (2014). Childhood sleep and lifelong mortality risk. *Health Psychology*, 33, 1195–1203.
137. Dumont, M., Leclerc, D., & McKinnon, S. (2009). Consequences of part-time work on the academic and psychosocial adaptation of adolescents. *Canadian Journal of School Psychology*, 24, 58–75.
138. Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, 14, 4–58.
139. Dunn, E. W., Aknin, L. B., & Norton, M. I. (2008). Spending money on others promotes happiness. *Science*, 319, 1687–1688.
140. Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, 105, 399–412.
141. Earp, B. D., Everett, J. A., Madva, E. N., & Hamlin, J. K. (2014). Out, damned spot: Can the “Macbeth Effect” be replicated? *Basic and Applied Social Psychology*, 36, 91–98.
142. Eich, E. (2014). Business not as usual [Editorial]. *Psychological Science*, 25, 3–6.
143. Ekman, P., & Friesen, W. V. (1978). *Facial action coding system: A technique for the measurement of facial movement*. Palo Alto, CA: Consulting Psychologists Press.
144. Epstein, Y. M., Suedfeld, P., & Silverstein, S. J. (1973). The experimental contract: Subjects' expectations of and reactions to some behaviors of experimenters. *American Psychologist*, 28, 212–221.
145. Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review*, 25, 5–34.

146. Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, 90, 891–904.
147. Farzan, R., & Kraut, R. E. (2013). Wikipedia classroom experiment: Bidirectional benefits of students' engagement in online production communities. *CHI'13, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 783–792.
148. Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175–191.
149. Feldman-Barrett, L., & Barrett, D. J. (2001). Computerized experience-sampling: How technology facilitates the study of conscious experience. *Social Science Computer Review*, 19, 175–185.
150. Felten, E. W. (2014, June 6). Facebook's emotional manipulation study: When ethical worlds collide [Weblog]. *Huffington Post*. Retrieved from http://www.huffingtonpost.com/edward-w-felten/facebook-emotional-manip_b_5545567.html
151. Ferguson, R., Robidoux, S., & Besner, D. (2009). Reading aloud: Evidence for contextual control over lexical activation. *Journal of Experimental Psychology: Human Perception and Performance*, 35, 499–507.
152. Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the *Journal of Applied Psychology*: Little evidence of reform. *Educational and Psychological Measurement*, 61, 181–210.
153. Finkel, E. J., Eastwick, P. W., & Matthews, J. (2007). Speed-dating as an invaluable tool for studying romantic attraction: A methodological primer. *Personal Relationships*, 14, 149–166.
154. Fisher, C. B. (2010). Enhancing HIV vaccine trial consent preparedness among street drug users. *Journal of Empirical Research*

on Human Research Ethics: An International Journal, 5, 65–80.

155. Fiske, S. T. (2004). Mind the gap: In praise of informal sources of formal theory. *Personality and Social Psychology Review*, 8, 132–137.
156. Flavell, J. H. (1996). Piaget's legacy. *Psychological Science*, 7, 200–203.
157. Foster, R. G., & Roenneberg, T. (2008). Human responses to the geophysical daily, annual and lunar cycles. *Current Biology*, 18, R784–R794.
158. Fournier, J. C., DeRubeis, R. J., Hollon, S. D., Dimidjian, S., Amsterdam, J. D., Shelton, R. C., & Fawcett, J. (2010). Antidepressant drug effects and depression severity: A patient-level meta-analysis. *Journal of the American Medical Association*, 303, 47–53.
159. Fowler, F. J., Jr. (1984). *Survey research methods*. Newbury Park, CA: Sage.
160. Fraley, R. C., & Marks, M. J. (2007). The null hypothesis significance testing debate and its implications for personality research. In R. W. Robins, R. C. Fraley, and R. Krueger (Eds.), *Handbook of Research Methods in Personality Psychology* (pp. 149–169). New York: Guilford Press.
161. Fraser, K., Huffman, J., Ma, I., Sobczak, M., McIlwrick, J., Wright, B., & McLaughlin, K. (2014). The emotional and cognitive impact of unexpected simulated patient death: A randomized control trial. *CHEST*, 145, 958–963.
162. Freedman, J. L. (1969). Role-playing: Psychology by consensus. *Journal of Personality and Social Psychology*, 13, 107–114.
163. Frick, R. W. (1995). Accepting the null hypothesis. *Memory and Cognition*, 25, 132–138.

164. Friedman, H. S., Tucker, J. S., Schwartz, J. E., Tomlinson-Keasy, C., Martin, L. R., Wingard, D. L., & Criqui, M. H. (1995). Psychosocial and behavioral predictors of longevity: The aging and death of the "Termites." *American Psychologist*, 50, 69–78. Page RE-6
165. Gallup, G. G., & Suarez, S. D. (1985). Alternatives to the use of animals in psychological research. *American Psychologist*, 40, 1104–1111.
166. Gamble, T., & Walker, I. (2016). Wearing a bicycle helmet can increase risk taking and sensation seeking in adults. *Psychological Science*, 27, 289–294.
167. Gardner, G. T. (1978). Effects of federal human subjects regulations on data obtained in environmental stressor research. *Journal of Personality and Social Psychology*, 36, 628–624.
168. Gardner, L. E. (1988). A relatively painless method of introduction to the psychological literature search. In M. E. Ware & C. L. Brewer (Eds.), *Handbook for teaching statistics and research methods*. Hillsdale, NJ: Erlbaum.
169. Gaucher, D., Friesen, J., & Kay, A. C. (2011). Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of Personality and Social Psychology*, 101, 109–128.
170. Gaudreau, P., Miranda, D., & Gareau, A. (2014). Canadian university students in wireless classrooms: What do they do on their laptops and does it really matter? *Computers and Education*, 70, 245–255.
171. Gauthier, C. (2004). Overview and analysis of animal use in North America. *ALTA: Alternatives to Laboratory Animals*, 32, 275–285.
172. Gee, C. J., & Leith, L. M. (2007). Aggressive behavior in professional ice hockey: A cross-cultural comparison of North American and European born NHL players. *Psychology of Sport and Exercise*, 8, 567–583.

173. Gelfand, M. J., & Diener, E. (2010). Culture and psychological science: Introduction to the special section. *Perspectives on Psychological Science*, 5, 390.
174. Gervais, W. M. (2011). Finding the faithless: Perceived atheist prevalence reduces anti-atheist prejudice. *Personality and Social Psychology Bulletin*, 37, 543–556.
175. Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, 8, 53–96.
176. Gilovich, T. (1991). *How we know what isn't so: The fallibility of human reason in everyday life*. New York: Free Press.
177. Giltrow, J., Gooding, R., Burgoyne, D., & Sawatsky, M. (2014). *Academic Writing: An Introduction* (3rd ed.). Peterborough, ON: Broadview.
178. Glantz, L. H., Annas, G. J., Grodin, M. A., Mariner, W. K. (2001). Research in developing countries: Taking “benefit” seriously. In Teays, W., & Purdy, L. (Eds.), *Bioethics, justice, and health care* (pp. 261–267). Belmont, CA: Wadsworth.
179. Gonzalez, A. Q., & Koestner, R. (2005). Parental preference for sex of newborn as reflected in positive affect in birth announcements. *Sex Roles*, 52, 407–411.
180. Goodman, S. (2008). A dirty dozen: Twelve *P*-value misconceptions. *Seminars in Hematology*, 45, 135–140.
181. Goodstein, D. (2010). *On fact and fraud: Cautionary tales from the front lines of science*. Princeton, NJ: Princeton University Press.
182. Goodstein, D. (2011). How science works. In Committee on Science, Technology, and Law Policy and Global Affairs, *Reference Manual on*

Scientific Evidence (3rd ed.) (pp. 37–54). Washington, DC: The National Academies Press. Retrieved from <http://www.fjc.gov/>

183. Gosling, S. D., & Johnson, J. A. (Eds.). (2010). *Advanced methods for behavioral research on the Internet*. Washington, DC: American Psychological Association.
184. Gosling, S. D., Sandy, C. J., & Potter, J. (2010). Personalities of self-identified “dog people” and “cat people.” *Anthrozoös*, 23, 213–222.
185. Gosling, S. D., Vazire, S., Srivastava, S., & John, O. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about Internet questionnaires. *American Psychologist*, 59, 93–104.
186. Graesser, A. C., Kennedy, T., Wiemer-Hastings, P., & Ottati, V. (1999). The use of computational cognitive methods to improve questions on surveys and questionnaires. In M. G. Sirkin, D. J. Hermann, S. Schechter, N. Schwarz, J. M. Tanur, & R. Tourangeau (Eds.), *Cognition and survey methods research* (pp. 199–216). New York: Wiley.
187. Graham, K., Bernards, S., Osgood, D. W., Parks, M., Abbey, A., Felson, R. B., . . . Wells, S. (2013). Apparent motives for aggression in the social context of the bar. *Psychology of Violence*, 3, 218–232.
188. Graham, K., Tremblay, P. F., Wells, S., Pernanen, K., Purcell, J., & Jolley, J. (2006). Harm, intent, and the nature of aggressive behaviour: Measuring naturally occurring aggression in barroom settings. *Assessment*, 13, 280–296.
189. Grant, D. A. (1948). The latin square principle in the design and analysis of psychological experiments. *Psychological Bulletin*, 45, 427–442.
190. Grant, T., Furlano, R., Hall, L., & Kelley, E. (2018). Criminal responsibility in autism spectrum disorder: A critical review

examining empathy and moral reasoning. *Canadian Psychology/Psychologie canadienne*, 59, 65–75.

191. Green, J., & Wallaf, C. (1981). *Ethnography and language in educational settings*. New York: Ablex.
192. Green, R. J., Sandall, J. C., & Phelps, C. (2005). Effect of experimenter attire and sex on participant productivity. *Social Behavior and Personality*, 33, 125–132.
193. Greenfield, D. N. (1999). *Nature of Internet addiction: Psychological factors in compulsive Internet use*. Paper presented at the meeting of the American Psychological Association, Boston, MA.
194. Greenland, S., & Morgenstern, H. (2001). Confounding in health research. *Annual Review of Public Health*, 22, 189–212.
195. Greenwald, A. G. (1976). Within-subjects designs: To use or not to use? *Psychological Bulletin*, 83, 314–320.
196. Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: Univariate and multivariate applications* (2nd ed.). New York: Routledge.
197. Gross, A. E., & Fleming, I. (1982). Twenty years of deception in social psychology. *Personality and Social Psychology Bulletin*, 8, 402–408.
198. Guay, J.-P., Ruscio, J., Knight, R. A., & Hare, R. D. (2007). A taxometric analysis of the latent structure of psychopathy: Evidence for dimensionality. *Journal of Abnormal Psychology*, 116, 701–716.
199. Hamamura, T., Heine, S. J., & Paulhus, D. L. (2008). Cultural differences in response styles: The role of dialectical thinking. *Personality and Individual Differences*, 44, 932–942. Page RE-7
200. Hammond, D., Fong, G. T., Borland, R., Cummings, K. M., McNeill, A., & Driezen, P. (2007). Text and graphic warnings on cigarette

packages: Findings from the International Tobacco Control Four Country Study. *American Journal of Preventive Medicine*, 32, 202–209.

201. Haney, C., & Zimbardo, P. G. (1998). The past and future of U.S. prison policy: Twenty-five years after the Stanford Prison Experiment. *American Psychologist*, 53, 709–727.
202. Hanson, M. D., & Chen, E. (2010). Daily stress, cortisol, and sleep: The moderating role of childhood psychosocial environments. *Health Psychology*, 29, 394–402.
203. Hare, R. D. (1991). *The Hare Psychopathy Checklist-Revised (PCL-R)*. Toronto: Multi-Health Systems.
204. Hare, R. D., Harpur, T. J., & Hemphill, J. D. (1989). Scoring Pamphlet for the Self-Report Psychopathy Scale: SRP-II. Unpublished manuscript, Simon Fraser University, Vancouver, British Columbia, Canada.
205. Harlow, L. L., & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological Methods*, 21, 447–457.
206. Harmon, L. W., DeWitt, D. W., Campbell, D. P., & Hansen, J. I. C. (1994). *Strong interest inventory: Applications and technical guide: form T317 of the Strong vocational interest blanks*. Palo Alto, CA: Stanford University Press.
207. Harris, C. M., & Cameron, S. L. (2010). Displacing *Wikipedia*: Information literacy for first-year students. In D. S. Dunn, B. C. Beins, M. A. McCarthy, & G. W. Hill, IV (Eds.), *Best practices for teaching beginnings and endings in the psychology major* (pp. 125–136). New York: Oxford.
208. Harris, C. R., Coburn, N., Rohrer, D., & Pashler, H. (2013). Two failures to replicate high-performance-goal priming effects. *PLoS ONE*, 8, e72467.

209. Harrison, N., Mordell, S., Roesch, R., & Watt, K. (2015). Patients with mental health issues in the emergency department: The relationship between coercion and perceptions of being helped, psychologically hurt, and physically harmed. *International Journal of Forensic Mental Health*, 14, 161–171.
210. Hart, S. D., & Hare, R. D. (1989). Discriminant validity of the Psychopathy Checklist in a forensic psychiatric population. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, 1, 211–218.
211. Hathaway, S. R., McKinley, J. C., & MMPI Restandardization Committee. (1989). *MMPI-2: Minnesota Multiphasic Personality Inventory-2: manual for administration and scoring*. Minneapolis, MN: University of Minnesota Press.
212. Hawking, S. W. (1988). *A brief history of time: From the big bang to black holes*. New York: Bantam Books.
213. Hayashi, K., Wood, E., Wiebe, L., Qi, J., & Kerr, T. (2010). An external evaluation of a peer-run outreach-based syringe exchange in Vancouver, Canada. *International Journal of Drug Policy*, 21, 418–421.
214. Heine, S. J., Foster, J. A. B., & Spina, R. (2009). Do birds of a feather universally flock together? Cultural variation in the similarity-dattraction effect. *Asian Journal of Social Psychology*, 12, 247–258.
215. Heine, S. J., Lehman, D. R., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales? The reference-group effect. *Journal of Personality and Social Psychology*, 82, 903–918.
216. Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist*, 58, 78–80.
217. Henrich, J., Heine, S. J., & Norenzayan, A. (2010a). The weirdest people in the world? [Target Article]. *Behavioral and Brain Sciences*,

33, 61–83.

218. Henrich, J., Heine, S. J., & Norenzayan, A. (2010b). Beyond WEIRD: Towards a broad-based behavioural science. [Response]. *Behavioral and Brain Sciences*, 33, 111–135.
219. Henry, B., & Pulcino, R. (2009). Individual difference and study-specific characteristics influencing attitudes about the use of animals in medical research. *Society and Animals*, 17, 305–324.
220. Hertwig, R., & Ortmann, A. (2008). Deception in experiments: Revisiting the arguments in its defense. *Ethics & Behavior*, 18, 59–92.
221. Hill, L. (1990). Effort and reward in college: A replication of some puzzling findings. In J. W. Neuliep (Ed.), *Handbook of replication in the behavioral and social sciences* [Special issue]. *Journal of Social Behavior and Personality*, 5, 151–161.
222. Hockey, G. R., & Earle, F. (2006). Control over the scheduling of simulated office work reduces the impact of workload on mental fatigue and task performance. *Journal of Experimental Psychology: Applied*, 12, 50–65.
223. Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E. J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, 21, 1157–1164.
224. Hogarth, R. M., Portell, M., & Cuxart, A. (2007). What risks do people perceive in everyday life? A perspective gained from the Experience Sampling Method (ESM). *Risk Analysis*, 27, 1427–1439.
225. Hood, T. C., & Back, K. W. (1971). Self-disclosure and the volunteer: A source of bias in laboratory experiments. *Journal of Personality and Social Psychology*, 17, 130–136.
226. Howell, R. T., Rodzon, K. S., Kurai, M., & Sanchez, A. H. (2010). A validation of well-being and happiness surveys for administration via the Internet. *Behavior Research Methods*, 42, 775–784.

227. Huberty, C. J. (2003). Multiple correlation versus multiple regression. *Educational and Psychological Measurement*, 63, 271–278.
228. Interagency Advisory Panel on Research Ethics (PRE). (2009). *Introductory Tutorial for the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (TCPS)*. Retrieved from <http://www.pre.ethics.gc.ca/english/tutorial/welcome.cfm>
229. Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, 2, 0696–0701.
230. John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23, 524–532.
231. Jones, R., & Cooper, J. (1971). Mediation of experimenter effects. *Journal of Personality and Social Psychology*, 20, 70–74.
232. Judd, C. M., Smith, E. R., & Kidder, L. H. (1991). *Research methods in social relations* (6th ed.). Ft. Worth, TX: Holt, Rinehart & Winston.
233. Kail, R. V. (2015). *Scientific writing in psychology: Lessons in clarity and style*. Thousand Oaks, CA: SAGE. Page RE-8
234. Kaufmann, H. (1967). The price of obedience and the price of knowledge. *American Psychologist*, 22, 321–322.
235. Kazdin, A. E. (1995). Preparing and evaluating research reports. *Psychological Assessment*, 7, 228–237.
236. Kazdin, A. E. (2001). *Behavior modification in applied settings* (6th ed.). Belmont, CA: Wadsworth.
237. Kelman, H. C. (1967). Human use of human subjects: The problem of deception in social psychological experiments. *Psychological Bulletin*, 67, 1–11.
238. Kenny, D. A. (1979). *Correlation and causality*. New York: Wiley.

239. Kifer, Y., Heller, D., Perunovic, W. Q. E., & Galinsky, A. D. (2013). The good life of the powerful: The experience of power and authenticity enhances subjective well-being. *Psychological Science*, 24, 280–288.
240. Kintz, N. L., Delprato, D. J., Mettee, D. R., Persons, C. E., & Schappe, R. H. (1965). The experimenter effect. *Psychological Bulletin*, 63, 223–232.
241. Kirsch, I., Deacon, B. J., Huedo-Medina, T. B., Scoboria, A., Moore, T. J., & Johnson, B. T. (2008). Initial severity and antidepressant benefits: A meta-analysis of data submitted to the Food and Drug Administration. *PLoS Medicine*, 5, 260–268.
242. Kirschbaum, C., Pirke, K.-M., & Hellhammer, D. H. (1993). The ‘Trier Social Stress Test’: A tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, 28, 76–81.
243. Klassen, R. M., & Chiu, M. M. (2010). Effects of teachers’ self-efficacy and job satisfaction: Teacher gender, years of experience, and job stress. *Journal of Educational Psychology*, 102, 741–759.
244. Klatzky, R. L. (2009). Giving psychological science away: The role of applications courses. *Perspectives on Psychological Science*, 4, 522–530.
245. Klaver, J. R., Lee, Z., & Hart, S. D. (2007). Psychopathy and nonverbal indicators of deception in offenders. *Law and Human Behavior*, 31, 337–351.
246. Klein, O., Doyen, S., Leys, C., Magalhães de Saldanha da Gama, P. A., Miller, S., Questienne, L., & Cleeremans, A. (2012). Low hopes, high expectations: Expectancy effects and the replicability of behavioural experiments. *Perspectives on Psychological Science*, 7, 572–584.

247. Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., . . . & Batra, R. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1, 443–490.
248. Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford Press.
249. Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences* (2nd ed.). Washington, DC: APA Books.
250. Knight, S., Vrij, A., Bard, K., & Brandon, D. (2009). Science versus human welfare? Understanding attitudes toward animal use. *Journal of Social Issues*, 65, 463–483.
251. Knoll, A. D., & MacLennan, R. N. (2017). Prevalence and correlates of depression in Canada: Findings from the Canadian Community Health Survey. *Canadian Psychology/Psychologie canadienne*, 58, 116–123.
252. Koocher, G. P. (1977). Bathroom behavior and human dignity. *Journal of Personality and Social Psychology*, 35, 120–121.
253. Koocher, G. P. (2009). Ethics and the invisible psychologist. *Psychological Services*, 6, 97–107.
254. Korn, J. H. (1997). *Illusions of reality: A history of deception in social psychology*. Albany: State University of New York Press.
255. Koss, M. P. (1992). The underdetection of rape: Methodological choices influence incident estimates. *Journal of Social Issues*, 48, 61–75.
256. Kotovych, M., Dixon, P., Bortolussi, M., & Holden, M. (2011). Textual determinants of a component of literary identification. *Scientific Study of Literature*, 1, 260–291.

257. Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*, 8788–8790.
258. Krätzig, G. P., & Arbuthnott, K. D. (2006). Perceptual learning style and learning proficiency: A test of the hypothesis. *Journal of Educational Psychology*, *98*, 238–246.
259. Kraut, R., Olson, J., Banaji, M., Bruckman, A., Cohen, J., & Couper, M. (2004). Psychological research online: Report of Board of Scientific Affairs Advisory Group on the Conduct of Research on the Internet. *American Psychologist*, *59*, 105–117.
260. Krawczyk, M. (2011). What brings your subjects to the lab? A field experiment. *Experimental Economics*, *14*, 482–489.
261. Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*, 537–567.
262. Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PLoS ONE*, *9*, e105825.
263. Kwan, D., Craver, C. F., Green, L., Myerson, J., & Rosenbaum, R. S. (2013). Dissociations in future thinking following hippocampal damage: Evidence from discounting and time perspective in episodic amnesia. *Journal of Experimental Psychology: General*, *142*, 1355–1369.
264. Lamothe, M., Boujut, E., Zenasni, F., & Sultan, S. (2014). To be or not to be empathic: the combined role of empathic concern and perspective taking in understanding burnout in general practice. *BMC Family Practice*, *15*, 15.
265. Lana, R. E. (1969). Pretest sensitization. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifacts in behavioral research*. New York: Academic Press.

266. Laney, C., Kaasa, S. O., Morris, E. K., Berkowitz, S. R., Bernstein, D. M., & Loftus, E. F. (2008). The Red Herring technique: A methodological response to the problem of demand characteristics. *Psychological Research*, 72, 362–375.
267. Larsen, H., Overbeek, G., Granic, I., & Engels, R. C. (2012). The strong effect of other people's drinking: Two experimental observation studies in a real bar. *American Journal on Addictions*, 21, 168–175.
268. Laverty, W. H., & Kelly, I. W. (1998). Cyclical calendar and lunar patterns in automobile property accidents and injury accidents. *Perceptual and Motor Skills*, 86, 299–302.
269. LeBel, E. P., & Campbell, L. (2009). Implicit partner affect, relationship satisfaction, and the prediction of romantic breakup. *Journal of Experimental Social Psychology*, 45, 1291–1294.
270. LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology*, 15, 371–379. Page RE-9
271. LeBel, E. P., Borsboom, D., Giner-Sorolla, R., Hasselman, F., Peters, K. R., Ratliff, K. A., & Smith, C. T. (2013). PsychDisclosure.org: Grassroots support for reforming reporting standards in psychology. *Perspectives on Psychological Science*, 8, 424–432.
272. Levelt Committee, Noort Committee, Drenth Committee (2012, November 28). *Flawed science: The fraudulent research practices of social psychologist Diederik Stapel* [English translation]. Retrieved from https://www.commissielevelt.nl/wp-content/uploads_per_blog/commissielevelt/2013/01/finalreportLevelt1.pdf
273. Levine, D. G., & Ducharme, J. M. (2013). The effects of a teacher-child play intervention on classroom compliance in young children in child care settings. *Journal of Behavioral Education*, 22, 50–65.

274. Levine, R. V. (1990). The pace of life. *American Scientist*, 78, 450–459.
275. Lewis, S. C., Zamith, R., & Hermida, A. (2013). Content analysis in an era of Big Data: A hybrid approach to computational and manual methods. *Journal of Broadcasting & Electronic Media*, 57, 34–52.
276. Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22(140), 5–55.
277. Lilienfeld, S. O., Ritschel, L. A., Lynn, S. J., Cautin, R. L., & Latzman, R. D. (2013). Why many clinical psychologists are resistant to evidence-based practice: Root causes and constructive remedies. *Clinical Psychology Review*, 33, 883–900.
278. Lima, J., McCabe-Bennett, H., & Antony, M. M. (2018). Treatment of storm fears using virtual reality and progressive muscle relaxation. *Behavioural & Cognitive Psychotherapy*, 46, 251–256.
279. Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics*, 15, 22–25.
280. Lindsay, R. C., & Wells, G. L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. *Journal of Applied Psychology*, 70, 556–564.
281. Linden, W., Talbot Ellis, A., & Millman, R. (2010). Deception in stress reactivity and recovery research. *International Journal of Psychophysiology*, 75, 33–38.
282. Lönnqvist, J-E., Paunonen, S., Verkasalo, M., Leikas, S., Tuulio-Henriksson, A., & Lönnqvist, J. (2007). Personality characteristics of research volunteers. *European Journal of Personality*, 21, 1017–1030.
283. Luria, A. R. (1968). *The mind of a mnemonist*. New York: Basic Books.

284. Lynn, M., & McCall, M. (2009). Techniques for increasing servers' tips. *Cornell Hospitality Quarterly*, 50, 198–208.
285. Lynn, M., & Sturman, M. J. (2010). Tipping and service quality: A within-subjects analysis. *Journal of Tourism and Hospitality Research*, 34, 269–275.
286. MacDonald, G., & Borsook, T. K. (2010). Attachment avoidance and feelings of connection in social interaction. *Journal of Experimental Social Psychology*, 46, 1122–1125.
287. Madigan, R., Johnson, S., & Linton, P. (1995). The language of psychology: APA style as epistemology. *American Psychologist*, 50, 428–436.
288. Marjanovic, Z., & Holden, R. R. (2019). Differentiating conscientious from indiscriminate responders in existing NEO-Five Factor Inventory-3 data. *Journal of Research in Personality*, 81, 127–137.
289. Marjanovic, Z., Struthers, C. W., Cribbie, R. A., & Greenglass, E. R. (2014). The Conscientious Responders Scale: A new tool for discriminating between conscientious and random responders. *SAGE Open*, 4, 1–10.
290. Markowitz, A., Blaszkiewicz, K., Montag, C., Switala, C., & Schlaepfer, T. E. (2014). Psycho-informatics: Big data shaping modern psychometrics. *Medical Hypotheses*, 82, 405–411.
291. Masson, M. E. J., & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology*, 57, 203–220.
292. Matsumoto, D. (1994). *Cultural influences on research methods and statistics*. Belmont, CA: Brooks/Cole.
293. Mayhew, D. R., Simpson, H. M., Wood, K. M., Lonero, L., Clinton, K. M., & Johnson, A. G. (2011). On-road and simulated driving:

- Concurrent and discriminant validation. *Journal of Safety Research*, 42, 267–275.
294. McCutcheon, L. E. (2000). Another failure to generalize the Mozart effect. *Psychological Reports*, 87, 325–330.
295. McFerran, B., Dahl, D. W., Fitzsimons, G. J., & Morales, A. C. (2010). I'll have what she's having: Effects of social influence and body type on the food choices of others. *Journal of Consumer Research*, 36, 915–929.
296. McGuigan, F. J. (1963). The experimenter: A neglected stimulus. *Psychological Bulletin*, 60, 421–428.
297. McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, 22, 276–282.
298. McNeill, P. M. (1993). *The ethics and politics of human experimentation*. New York, NY: Cambridge University Press.
299. McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, 73(Suppl. 1), 235–245.
300. Mehdizadeh, S. (2010). Self-presentation 2.0: Narcissism and self-esteem on Facebook. *Cyberpsychology, Behavior, and Social Networking*, 13, 357–364.
301. Mehl, M. R., Vazire, S., Holleran, S. E., & Clark, C. S. (2010). Eavesdropping on happiness: Well-being is related to having less small talk and more substantive conversations. *Psychological Science*, 21, 539–541.
302. Middlemist, R. D., Knowles, E. S., & Matter, C. F. (1976). Personal space invasion in the lavatory: Suggestive evidence for arousal. *Journal of Personality and Social Psychology*, 33, 541–546.

303. Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., . . . Van der Laan, M. (2014). Promoting transparency in social science research. *Science*, 343(6166), 30–31.
304. Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, 67, 371–378.
305. Milgram, S. (1964). Group pressure and action against a person. *Journal of Abnormal and Social Psychology*, 69, 137–143.
306. Milgram, S. (1965). Some conditions of obedience and disobedience to authority. *Human Relations*, 18, 57–76.
307. Miller, A. G. (1972). Role-playing: An alternative to deception? *American Psychologist*, 27, 623–636. Page RE-10
308. Miller, A. G. (1986). *The obedience experiments: A case study of controversy in social science*. New York: Praeger.
309. Miller, G. A. (1969). Psychology as a means of promoting human welfare. *American Psychologist*, 24, 1063–1075.
310. Miller, J. G. (1999). Cultural psychology: Implications for basic psychological theory. *Psychological Science*, 10, 85–91.
311. Miller, N. E. (1985). The value of behavioral research on animals. *American Psychologist*, 40, 423–440.
312. Mitchell, G. (2012). Revisiting truth or triviality: The external validity of research in the psychological laboratory. *Perspectives on Psychological Science*, 7, 109–117.
313. Montee, B. B., Miltenberger, R. G., & Wittrock, D. (1995). An experimental analysis of facilitated communication. *Journal of Applied Behavior Analysis*, 28, 189–200.
314. Morgan, D. L. & Morgan, R. K. (2001). Single-participant research design: Bringing science to managed care. *American Psychologist*, 56, 119–127.

315. Mosby, I. (2013). Administering colonial science: Nutrition research and human biomedical experimentation in Aboriginal communities and residential schools, 1942–1952. *Histoire sociale/Social history*, 46(91), 145–172.
316. Moss, E. L. & von Ranson, K. M. (2006). An experimental investigation of recruitment bias in eating pathology research. *International Journal of Eating Disorders*, 39, 256–259.
317. Mostert, M. (2010). Facilitated communication and its legitimacy—Twenty-first century developments. *Exceptionality*, 18, 31–41.
318. Moulden, H. M., Firestone, P., Kingston, D. A., & Wexler, A. F. (2010). A description of sexual offending committed by Canadian teachers. *Journal of Child Sexual Abuse*, 19, 403–418.
319. Murphy, F. C., & Klein, R. M. (1998). The effects of nicotine on spatial and non-spatial expectancies in a covert orienting task. *Neuropsychologia*, 36, 1103–1114.
320. Murray, B. (2002). Research fraud needn't happen at all. *NAPA Monitor*, 33(2). Retrieved from <http://www.apa.org/monitor/feb02/fraud.html>
321. Mychasiuk, R., & Benzies, K. (2011). Facebook: An effective tool for participant retention in longitudinal researcher. *Child: Care, Health and Development*, 38, 753–756.
322. Nathan, R. (2005). *My freshman year: What a professor learned by becoming a student*. Ithaca, NY: Cornell University Press.
323. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979, April 18). *The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research*. Retrieved from <http://ohsr.od.nih.gov/guidelines/belmont.html>

324. Nichols, A. L., & Maner, J. K. (2008). The good-subject effect: Investigating participant demand characteristics. *The Journal of General Psychology*, 135, 151–165.
325. Nicol, A. A. M., & Pexman, P. M. (2010). *Displaying your findings: A practical guide for creating figures, posters, and presentations* (6th ed.). Washington, DC: American Psychological Association.
326. Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall.
327. Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 231–259.
328. Norman, M. (2012). Saturday night's alright for tweeting: Cultural citizenship, collective discussion, and the new media consumption/production of *Hockey Day in Canada*. *Sociology of Sport Journal*, 29, 306–324.
329. Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., . . . & Contestabile, M. (2015). Promoting an open research culture. *Science*, 348(6242), 1422–1425.
330. Nowlis, S. M., Kahn, B. E., & Dhar, R. (2002). Coping with ambivalence: The effect of removing a neutral option on consumer attitude and preference judgments. *Journal of Consumer Research*, 29, 319–334.
331. Nunes, F. (1998). *Portuguese-Canadians from sea to sea: A national needs assessment*. Toronto: Portuguese-Canadian National Congress.
332. Open Science Collaboration. (2013). *The Reproducibility Project: A model of large-scale collaboration for empirical research on reproducibility*. Retrieved from <http://ssrn.com/abstract=2195999>
333. Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.

334. Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, 17, 776–783.
335. Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois Press.
336. Osumi, T., & Ohira, H. (2010). The positive side of psychopathy: Emotional detachment in psychopathy and rational decision-making in the ultimatum game. *Personality and Individual Differences*, 49, 451–456.
337. Ozdemir, A. (2008). Shopping malls: Measuring interpersonal distance under changing conditions and across cultures. *Field Methods*, 20, 226–248.
338. Pashler, H., Coburn, N., & Harris, C. R. (2012). Priming of social distance? Failure to replicate effects on social and food judgments. *PLoS ONE*, 7, e42510.
339. Patterson, M. L., Lammers, V. M., & Tubbs, M. E. (2014). Busy signal: Effects of mobile device usage on pedestrian encounters. *Journal of Nonverbal Behavior*, 38, 313–324.
340. Paulhus, D. L., Harms, P. D., Bruce, M. N., & Lysy, D. C. (2003). The Over-Claiming Technique: Measuring self-enhancement independent of ability. *Journal of Personality and Social Psychology*, 84, 890–904.
341. Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). Linguistic Inquiry and Word Count: LIWC [Computer software]. Austin, TX: LIWC.net.
342. Pepitone, A., & Triandis, H. (1987). On the universality of social psychological theories. *Journal of Cross-Cultural Psychology*, 18, 471–499.
343. Pfungst, O. (1911). *Clever Hans (the horse of Mr. von Osten): A contribution to experimental, animal, and human psychology* (C. L.

Rahn, Trans.). New York: Holt, Rinehart & Winston. (Republished 1965.)Page RE-11

344. Philipp-Muller, A., & MacDonald, G. (2017). Avoidant individuals may have muted responses to social warmth after all: An attempted replication of MacDonald and Borsook (2010). *Journal of Experimental Social Psychology*, 70, 272–280.
345. Pietschnig, J., Voracek, M., & Formann, A. K. (2010). Mozart effect–Shmozart effect: A meta-analysis. *Intelligence*, 38, 314–323.
346. Plous, S. (1996a). Attitudes toward the use of animals in psychological research and education: Results from a national survey of psychologists. *American Psychologist*, 51, 1167–1180.
347. Plous, S. (1996b). Attitudes toward the use of animals in psychological research and education: Results from a national survey of psychology majors. *Psychological Science*, 7, 352–363.
348. Popper, K. (1968). *The logic of scientific discovery*. New York: Harper & Row.
349. Pozzulo, J. D., & Marciniak, S. (2006). Comparing identification procedures when the perpetrator has changed appearance. *Psychology, Crime & Law*, 12, 429–438.
350. Pozzulo, J. D., Crescini, C., & Panton, T. (2008). Does methodology matter in eyewitness identification research?: The effect of live versus video exposure on eyewitness identification accuracy. *International Journal of Law and Psychiatry*, 31, 430–437.
351. 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.
352. Prelec, D. (2004). A Bayesian Truth Serum for subjective data. *Science*, 306(5695), 462–466.

353. Provencher, V., Bier, N., Audet, T., & Gagnon, L. (2008). Errorless-based techniques can improve route finding in early Alzheimer's disease: A case study. *American Journal of Alzheimer's Disease & Other Dementias*, 23, 47–56.
354. Provost, M. P., Kormos, C., Kosakoski, G., & Quinsey, V. L. (2006). Sociosexuality in women and preference for facial masculinization and somatotype in men. *Archives of Sexual Behavior*, 35, 305–312.
355. Pruessner, M., Béchard-Evans, L., Boekstyn, L., Iyer, S. N., Pruessner, J. C., & Malla, A. K. (2013). Attenuated cortisol response to acute psychosocial stress in individuals at ultra-high risk for psychosis. *Schizophrenia Research*, 146, 79–86.
356. Prus, R., & Irini, S. (1980). *Hookers, rounders, and desk clerks: The social organization of the hotel community*. Toronto: Gage.
357. Public Prosecution Service of Canada (PPSC). (2011). Federal/Provincial/Territorial heads of Prosecutions Subcommittee on the Prevention of Wrongful Convictions. *The path to justice: Preventing wrongful convictions*. Retrieved from <http://www.ppsc-sppc.gc.ca/eng/pub/ptj-spj/index.html>
358. Quinlan, C. K., Taylor, T. L., & Fawcett, J. M. (2010). Directed forgetting: Comparing pictures and words. *Canadian Journal of Experimental Psychology*, 64, 41–46.
359. R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
360. *R. v. Berikoff*, BCSC 1024 (CanLII). (2000). Retrieved from <http://www.canlii.org/en/bc/bcsc/doc/2000/2000bcsc1024/2000bcsc1024.html>
361. *R. v. Lavallee*, 1 SCR 852 (CanLII). (1990). Retrieved from <http://www.canlii.org/en/ca/scc/doc/1990/1990canlii95/1990canlii95.html>

362. *R. v. Trochym*, 1 SCR 239 (CanLII). (2007). Retrieved from <http://www.canlii.org/en/ca/scc/doc/2007/2007scc6/2007scc6.html>
363. Rasch, B., & Born, J. (2013). About sleep's role in memory. *Physiological Reviews*, 93, 681–766.
364. Rauscher, F. H., & Shaw, G. L. (1998). Key components of the Mozart effect. *Perceptual and Motor Skills*, 86, 835–841.
365. Rauscher, F. H., Shaw, G. L., & Ky, K. N. (1993). Music and spatial task performance. *Nature*, 365, 611.
366. Rauscher, F. H., Shaw, G. L., & Ky, K. N. (1995). Listening to Mozart enhances spatial-temporal reasoning: Towards a neurophysiological basis. *Neuroscience Letters*, 185, 44–47.
367. Ravizza, S. M., Uitvlugt, M. G., & Fenn, K. M. (2017). Logged in and zoned out: How laptop Internet use relates to classroom learning. *Psychological Science*, 28, 171–180.
368. Rawn, C. D., & Vohs, K. D. (2011). People use self-control to risk personal harm: An intra-interpersonal dilemma. *Personality and Social Psychology Review*, 15, 267–289.
369. Reed, J. G., & Baxter, P. M. (2003). *Library use: A handbook for psychology* (3rd ed.). Washington, DC: American Psychological Association.
370. Reid, A. (2013, July 8). Angus Reid: What went wrong with the polls in British Columbia? *Maclean's*. Retrieved from <http://www.macleans.ca/news/canada/angus-reid-what-went-wrong-with-the-polls-in-british-columbia/>
371. Report of the Smeesters follow-up investigation committee. (2014). Retrieved from http://www.rsm.nl/fileadmin/Images_NEW/News_Images/2014/Report_Smeesters_follow-up_investigation_committee.final.pdf

372. Richard, F. D., Bond, C. F., Jr., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, 7, 331–363.
373. Richards, N. M., & King, J. H. (2014). Big data ethics. *Wake Forest Law Review*, 49, 393–432.
374. Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin*, 138, 353–387.
375. Ring, K., Wallston, K., & Corey, M. (1970). Mode of debriefing as a factor affecting subjective reaction to a Milgram-type obedience experiment: An ethical inquiry. *Representative Research in Social Psychology*, 1, 67–88.
376. Roberson, M. T., & Sundstrom, E. (1990). Questionnaire design, return rates, and response favorableness in an employee attitude questionnaire. *Journal of Applied Psychology*, 75, 354–357.
377. Robinson, J. P., Rusk, J. G., & Head, K. B. (1968). *Measures of political attitudes*. Ann Arbor, MI: Institute for Social Research.
378. Robinson, J. P., Shaver, P. R., & Wrightsman, L. S. (1991). *Measures of personality and social psychological attitudes* (Vol. 1). San Diego, CA: Academic Press. Page RE-12
379. Roehrs, T., Burduvali, E., Bonahoom, A., Drake, C., & Roth, T. (2003). Ethanol and sleep loss: A “dose” comparison of impairing effects. *Sleep*, 26, 981–985.
380. Rosenhan, D. (1973). On being sane in insane places. *Science*, 179(4070), 250–258.
381. Rosenthal, R. (1966). *Experimenter effects in behavior research*. New York: Appleton-Century-Crofts.

382. Rosenthal, R. (1967). Covert communication in the psychological experiment. *Psychological Bulletin*, 67, 356–367.
383. Rosenthal, R. (1991). *Meta-analytic procedures for social research* (rev. ed.). Newbury Park, CA: Sage.
384. Rosenthal, R. (2003). Covert communication in laboratories, classrooms, and the truly real word. *Current Directions in Psychological Science*, 12, 151–154.
385. Rosenthal, R., & Rosnow, R. L. (1975). *The volunteer subject*. New York: Wiley.
386. Rosnow, R. L., & Rosnow, M. (2012). *Writing papers in psychology* (9th ed.). Belmont, CA: Cengage Learning.
387. Ross, M., Xun, W. Q. E., & Wilson, A. E. (2002). Language and the bicultural self. *Personality and Social Psychology Bulletin*, 28, 1040–1050.
388. Rossi, P. H., Freeman, H. E., & Lipsey, M. W. (2004). *Evaluation: A systematic approach* (7th ed.). Thousand Oaks, CA: Sage.
389. Roszkowski, M. J., & Soven, M. (2010). Shifting gears: Consequences of including two negatively worded items in the middle of a positively worded questionnaire. *Assessment and Evaluation in Higher Education*, 35, 117–134.
390. Rothgerber, H., & Wolsiefer, K. (2014). A naturalistic study of stereotype threat in young female chess players. *Group Processes & Intergroup Relations*, 17, 79–90.
391. Rousseau, C., Benoit, M., Lacroix, L., & Gauthier, M.-F. (2009). Evaluation of a sandplay program for preschoolers in a multiethnic neighborhood. *Journal of Child Psychology and Psychiatry*, 50, 743–750.

392. Rozin, P. (2006). Domain denigration and process preference in academic psychology. *Perspectives on Psychological Science*, 1, 365–376.
393. Rozin, P. (2009). What kind of empirical research should we publish, fund, and reward? A different perspective. *Perspectives on Psychological Science*, 4, 435–439.
394. Rubin, Z. (1973). Designing honest experiments. *American Psychologist*, 28, 445–448.
395. Rubin, Z. (1975). Disclosing oneself to a stranger: Reciprocity and its limits. *Journal of Experimental Social Psychology*, 11, 233–260.
396. Ruggirello, C., & Mayer, C. (2010). Language development in a hearing and deaf twin with simultaneous bilateral cochlear implants. *Journal of Deaf Studies and Deaf Education*, 15, 274–286.
397. Runyan, W. M. (2006). Psychobiography and the psychology of science: Understanding relations between the life and work of individual psychologists. *Review of General Psychology*, 10, 147–162.
398. Russell, D., Peplau, L. A., & Cutrona, C. E. (1980). The revised UCLA Loneliness Scale: Concurrent and discriminant validity. *Journal of Personality and Social Psychology*, 39, 472–480.
399. Russell, W. M. S., & Burch, R. L. (1959). *The principles of humane experimental technique*. London: Methuen & Co.
400. Salkind, N. J. (2016). *Statistics for people who (think they) hate statistics: Using Microsoft Excel 2016*. Thousand Oaks, CA: Sage Publications.
401. Salkind, N. J., & Shaw, L. A. (2019). *Statistics for people who (think they) hate statistics: Using R*. Thousand Oaks, CA: Sage Publications.
402. Sarlon, E., Millier, A., Aballéa, S., & Tourni, M. (2014). Evaluation of different approaches for confounding in nonrandomized observational

data: A case-study of antipsychotics treatment. *Community Mental Health Journal*, 50, 711–720.

403. Scarapicchia, T. M. F., Sabiston, C. M., Andersen, R. E., & Garcia Bengoechea, E. (2013). The motivational effects of social contagion on exercise participation in young female adults. *Journal of Sport and Exercise Psychology*, 35, 563–575.
404. Schaie, K. W. (1986). Beyond calendar definitions of age, time, and cohort: The general developmental model revisited. *Developmental Review*, 6, 252–277.
405. Schellenberg, E. G. (2004). Music lessons enhance IQ. *Psychological Science*, 15, 511–514.
406. Schellenberg, E. G. (2006). Long-term positive associations between music lessons and IQ. *Journal of Educational Psychology*, 98, 457–468.
407. Schlenger, W. E., Caddell, J. M., Ebert, L., Jordan, B. K., Rourke, K. M., Wilson, D., . . . Kulk, R. A. (2002). Psychological reactions to terrorist attacks: Findings from the National Study of Americans' Reactions to September 11. *Journal of the American Medical Association*, 288, 581–588.
408. Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, 13, 90–100.
409. Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, 8, 350.
410. Schooler, J. W. (2014). Turning the lens of science on itself: Verbal overshadowing, replication, and metascience. *Perspectives on Psychological Science*, 9, 579–584.
411. Schur, E., Noonan, C., Polivy, J., Goldberg, J., & Buchwald, D. (2009). Genetic and environmental influences on restrained eating

- behavior. *International Journal of Eating Disorders*, 42, 765–772.
412. Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, 54, 93–105.
413. Scott, G. A., Saucier, D. M., & Lehmann, H. (2016). Contrasting the amnesic effects of temporary inactivation with lesions of the hippocampus on context memory. *Journal of Behavioral and Brain Science*, 6, 184–198.
414. 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.
415. Seifert, T., & Hedderson, C. (2009). Intrinsic motivation and flow in skateboarding: An ethnographic study. *Journal of Happiness Studies*, 11, 277–292.
416. Shadish, W. R. (2014). Statistical analyses of single-case designs: The shape of things to come. *Current Directions in Psychological Science*, 23, 139–146.
417. Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin. Page RE-13
418. Sharp, E. C., Pelletier, L. G., & Lévesque, C. (2006). The double-edged sword of rewards for participation in psychology experiments. *Canadian Journal of Behavioural Science*, 38, 269–277.
419. Sharpe, D., & Faye, C. (2009). A second look at debriefing practices: Madness in our method? *Ethics & Behavior*, 19, 432–447.
420. Shettleworth, S. J. (2009). The evolution of comparative cognition: Is the snark still a boojum? *Behavioural Processes*, 80, 210–217.
421. Sieber, J. E. (1992). *Planning ethically responsible research: A guide for students and internal review boards*. Newbury Park, CA: Sage.

422. Sieber, J. E., Iannuzzo, R., & Rodriguez, B. (1995). Deception methods in psychology: Have they changed in 23 years? *Ethics and Behavior*, 5, 67–85.
423. Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.
424. Silverman, I., & Margulis, S. (1973). Experiment title as a source of sampling bias in commonly used “subject-pool” procedures. *Canadian Psychologist/Psychologie Canadienne*, 14, 197–201.
425. 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.
426. Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2013). Life after p-hacking. Meeting of the Society for Personality and Social Psychology, New Orleans, LA, 17–19 January 2013. Available at SSRN: <http://ssrn.com/abstract=2205186> or <http://dx.doi.org/10.2139/ssrn.2205186>
427. Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, 9, 76–80.
428. Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered replication reports at perspectives on psychological science. *Perspectives on Psychological Science*, 9, 552–555.
429. Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science*, 24, 1875–1888.
430. Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file drawer. *Journal of Experimental Psychology: General*, 143, 534–547.

431. Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better P-Curves: Making P-Curve analysis more robust to errors, fraud, and ambitious P-hacking, a reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General*, 144, 1146–1152.
432. Sinclair, A. H., & Barense, M. D. (2018). Surprise and destabilize: prediction error influences episodic memory reconsolidation. *Learning & Memory*, 25, 369–381.
433. Skinner, B. F. (1953). *Science and human behavior*. New York: Macmillan.
434. Smart, R. (1966). Subject selection bias in psychological research. *Canadian Psychologist*, 7, 115–121.
435. Smith, C. P. (1983). Ethical issues: Research on deception, informed consent, and debriefing. In L. Wheeler & P. Shaver (Eds.), *Review of personality and social psychology* (Vol. 4). Newbury Park, CA: Sage.
436. Smith, R. J., Lingle, J. H., & Brock, T. C. (1979). Reactions to death as a function of perceived similarity to the deceased. *Omega*, 9, 125–138.
437. 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.
438. Snowden, D. A. (1997). Aging and Alzheimer's disease: Lessons from the Nun Study. *Gerontologist*, 37, 150–156.
439. Solomon, R. L. (1949). An extension of control group design. *Psychological Bulletin*, 46, 137–150.
440. Stanovich, K. E. (2013). *How to think straight about psychology* (10th ed.). Toronto: Pearson.
441. Statistics Canada. (2013). *NHS user guide: National Household Survey 2011*. Retrieved from <http://www12.statcan.gc.ca/nhs>

442. Steel, P. (2007). The nature of procrastination: A meta-analytic and theoretical review of quintessential self-regulatory failure. *Psychological Bulletin, 133*, 65–94.
443. Steele, K. M., Bass, K. E., & Crook, M. D. (1999). The mystery of the Mozart effect: Failure to replicate. *Psychological Science, 10*, 366–369.
444. Steinberg, L., & Dornbusch, S. M. (1991). Negative correlates of part-time employment during adolescence: Replication and elaboration. *Developmental Psychology, 27*, 304–313.
445. Sternberg, R. J., & Sternberg, K. (2010). *The psychologist's companion: a guide to writing scientific papers for students and researchers*. Cambridge University Press.
446. Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science, 7*, 670–688.
447. Strohmetz, D. B. (2008). Research artifacts and the social psychology of psychological experiments. *Social and Personality Psychology Compass, 2*, 861–877.
448. Suedfeld, P. (2010). The cognitive processing of politics and politicians: Archival studies of conceptual and integrative complexity. *Journal of Personality, 78*, 1669–1702.
449. Suedfeld P., & Jhangiani, R. (2009). Cognitive management in an enduring international rivalry: The case of India and Pakistan. *Political Psychology, 30*, 937–951.
450. Szabo, A., & Underwood, J. (2004). Cybercheats: Is information and communication technology fuelling academic dishonesty? *Active Learning in Higher Education, 5*, 180–199.

451. Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). New York: Pearson.
452. Tagliacollo, V. A., Volpato, G. L., & Pereira, A., Jr. (2010). Association of student position in classroom and school performance. *Educational Research*, 1, 198–201.
453. Tan, L. O., Hadjistavropoulos, T., & MacNab, Y. C. (2017). The Catastrophic Thoughts About Insomnia Scale (CTIS): Development and validation. *Cognitive Therapy and Research*, 41, 143–154.
454. Tao, D., Zhang, R., Lou, E., & Lalonde, R. N. (2018). The cultural shaping of career aspirations: Acculturation and Chinese biculturals' career identity styles. *Canadian Journal of Behavioural Science*, 50, 29–41.
455. Terman, L. M. (1925). *Genetic studies of genius: Vol. 1. Mental and physical traits of a thousand gifted children*. Stanford, CA: Stanford University Press. Page RE-14
456. Terman, L. M., & Oden, M. H. (1947). *Genetic studies of genius: Vol. 4. The gifted child grows up: Twenty-five years' follow-up of a superior group*. Stanford, CA: Stanford University Press.
457. Terman, L. M., & Oden, M. H. (1959). *Genetic studies of genius: Vol. 5. The gifted group in mid-life: Thirty-five years' follow-up of the superior child*. Stanford, CA: Stanford University Press.
458. The decline of smoking in Canada (2011, July 29). *CBC News*. Retrieved from <http://www.cbc.ca>
459. Theakston, J. A., Stewart, S. H., Dawson, M., Knowlden, S., & Lehman, D. R. (2004). Big five personality domains predict drinking motives. *Personality and Individual Differences*, 37, 971–984.
460. Thomas, G. V., & Blackman, D. (1992). The future of animal studies in psychology. *American Psychologist*, 47, 1679.

461. Thompson, W. F., Schellenberg, E. G., & Husain, G. (2001). Arousal, mood, and the Mozart effect. *Psychological Science*, 12, 248–251.
462. Tolin, D. F., Frost, R. O., Steketee, G., & Fitch, K. E. (2008). Family burden of compulsive hoarding: Results of an Internet survey. *Behaviour Research and Therapy*, 46, 334–344.
463. Tougas, F., Rinfret, N., Beaton, A. M., & de la Sablonnière, R. (2005). Policewomen acting in self-defense: Can psychological disengagement protect self-esteem from the negative outcomes of relative deprivation? *Journal of Personality and Social Psychology*, 88, 790–800.
464. Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, 133, 859–883.
465. Tracy, J. L., Robins, R. W., & Schriber, R. A. (2009). Development of a FACS-verified set of basic and self-conscious emotion expressions. *Emotion*, 9, 554–559.
466. Trafimow, D., & Marks, M. (2015). Editorial. M. *Basic and Applied Social Psychology*, 37, 1–2.
467. Trochim, W. M. (2000). *The research methods knowledge base* (2nd ed.). Cincinnati, OH: Atomic Dog Publishing.
468. Trochim, W. M. (2006). *The research methods knowledge base* (2nd ed.). Retrieved from <http://www.socialresearchmethods.net/kb/>
469. Uher, R., & Weaver, I. C. G. (2014). Epigenetic traces of childhood maltreatment in peripheral blood: A new strategy to explore gene-environment interactions. *British Journal of Psychiatry*, 204, 3–5.
470. Ullman, J. B. (2007) Structural equation modeling. In B. G. Tabachnick & L. S. Fidell (Eds.), *Using multivariate statistics* (5th ed.). New York: Allyn & Bacon.

471. U.S. Department of Justice. (1999). *Eyewitness evidence: A guide for law enforcement*. Retrieved from <http://www.ncjrs.org/pdffiles1/nij/178240.pdf>
472. Valtchanov, D., & Ellard, C. (2010). Physiological and affective responses to immersion in virtual reality: Effects of nature and urban settings. *Journal of CyberTherapy & Rehabilitation*, 3, 359–373.
473. Vanasse, A., Demers, M., Hemminki, A., & Courteau, J. (2006). Obesity in Canada: Where and how many? *International Journal of Obesity*, 30, 677–683.
474. Varao Sousa, T. L., Carriere, J. S. A., & Smilek, D. (2013). The way we encounter reading material influences how frequently we mind wander. *Frontiers in Psychology*, 4(892), 1–8.
475. Vazire, S. (2006). Informant reports: A cheap, fast, and easy method for personality assessment. *Journal of Research in Personality*, 40, 472–481.
476. Vazire, S., & Mehl, M. R. (2008). Knowing me, knowing you: The accuracy and unique predictive validity of self and other ratings of daily behavior. *Journal of Personality and Social Psychology*, 95, 1202–1216.
477. Verma, I. M. (2014). Editorial expression of concern: Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111, 10779.
478. Vicente, P., & Reis, E. (2010). Using questionnaire design to fight nonresponse bias in web surveys. *Social Science Computer Review*, 28, 251–267.
479. Wagstaff, G. F., MacVeigh, J., Boston, R., Scott, L., Brunas-Wagstaff, J., & Cole, J. (2003). Can laboratory findings on eyewitness testimony be generalized to the real world? An archival analysis of the influence

of violence, weapon presence, and age on eyewitness accuracy. *The Journal of Psychology*, 137, 17–28.

480. Wang, S.-Y., Parrila, R., & Cui, Y. (2013). Meta-analysis of social skills interventions of single-case research for individuals with autism spectrum disorders: Results from three-level HLM. *Journal of Autism and Developmental Disorders*, 43, 1701–1716.
481. Ward, M., Theule, J., & Cheung, K. (2016). Parent-child interaction therapy for child disruptive behaviour disorders: A meta-analysis. *Child & Youth Care Forum*, 45, 675–690.
482. Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: context, process, and purpose. *The American Statistician*, 70, 129–133.
483. Webb, E. J., Campbell, D. T., Schwartz, R. D., Sechrest, R., & Grove, J. B. (1981). *Nonreactive measures in the social sciences* (2nd ed.). Boston: Houghton Mifflin.
484. Weber, R. P. (1990). *Basic content analysis* (2nd ed.). Newbury Park, CA: Sage.
485. Wegner, D. M., Fuller, V. A., & Sparrow, B. (2003). Clever hands: Uncontrolled intelligence in facilitated communication. *Journal of Personality and Social Psychology*, 85, 5–19.
486. Weisberg, H. I. (2010). *Bias and causation: Models and judgment for valid comparisons*. Hoboken, NJ: John Wiley & Sons.
487. Wells, G. L. (2001). Police lineups: Data, theory, and policy. *Psychology, Public Policy, and Law*, 7, 791–801.
488. Wertz Garvin, A., & Damson, C. (2008). The effects of idealized fitness images on anxiety, depression and global mood states in college age males and females. *Journal of Health Psychology*, 13, 433–437.

489. Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLoS ONE*, 11, e0152719.
490. White, C. B., & Caird, J. K. (2010). The blind date: The effects of change blindness, passenger conversation and gender on looked-but-failed-to-see (LBFTS) errors. *Accident Analysis and Prevention*, 42, 1822–1830.
491. Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594–604. Page RE-15
492. Williams, K. M., Nathanson, C., & Paulhus, D. L. (2010). Identifying and profiling scholastic cheaters: Their personality, cognitive ability, and motivation. *Journal of Experimental Psychology: Applied*, 16, 293–307.
493. Windholz, G. (1997). Ivan P. Pavlov: An overview of his life and psychological work. *American Psychologist*, 52, 941–946.
494. Wintre, M. G., North, C., & Sugar, L. A. (2001). Psychologists' response to criticisms about research based on undergraduate participants: A developmental perspective. *Canadian Psychology*, 42, 216–225.
495. Wood, J. V., Perunovic, W. Q. E., & Lee, J. W. (2009). Positive self-statements: Power for some, peril for others. *Psychological Science*, 20, 860–866.
496. Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, 28, 186–191.
497. Wormith, J., Olver, M. E., Stevenson, H. E., & Girard, L. (2007). The long-term prediction of offender recidivism using diagnostic, personality, and risk/need approaches to offender assessment. *Psychological Services*, 4, 287–305.

498. Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, 44, 363–373.
499. Yarkoni, T. (2012). Psychoinformatics: New horizons at the interface of the psychological and computing sciences. *Current Directions in Psychological Science*, 21, 391–397.
500. Yarmey, A. D. (2003). Eyewitness identification: Guidelines and recommendations for identification procedures in the United States and in Canada. *Canadian Psychology*, 44, 181–189.
501. Yin, R. K. (1994). *Case study research: Design and methods*. Newbury Park, CA: Sage.
502. Yong, E. (2012, May 17). Replication studies: Bad copy. *Nature [News Feature]*, 485, 298–300.
503. Young, J. (2006, June 12). Wikipedia founder discourages academic use of his creation. *The Chronicle of Higher Education*. Retrieved from <http://chronicle.com/blogs/wiredcampus/wikipedia-founder-discourages-academic-use-of-his-creation/2305>
504. Yuille, J. C., Ternes, M., & Cooper, B. S. (2010). Expert testimony on laboratory witnesses. *Journal of Forensic Psychology Practice*, 10, 238–251.
505. Zerouali, Y., Jemel, B., & Godbout, R. (2010). The effects of early and late night partial sleep deprivation on automatic and selective attention: An ERP study. *Brain Research*, 1308, 87–99.
506. Zhong, C.-B., & DeVoe, S. E. (2010). You are how you eat: Fast food and impatience. *Psychological Science*, 21, 619–622.
507. Zimbardo, P. G. (1973). The psychological power and pathology of imprisonment. In E. Aronson & R. Helmreich (Eds.), *Social psychology*. New York: Van Nostrand.

508. Zimbardo, P. G. (2004). Does psychology make a significant difference in our lives? *American Psychologist*, 59, 339–351.

Sources

Chapter 12

Fig. 12.1, 12.2, 12.4, 12.5, 12.9: WORLD VALUES SURVEY Wave 6 2010-2014 OFFICIAL AGGREGATE v.20140429. World Values Survey Association (www.worldvaluessurvey.org). Aggregate File Producer: Asep/JDS, Madrid SPAIN. Retrieved from <http://www.worldvaluessurvey.org/>

Chapter 13

Fig 13.5: (1,4) Sana, F., Weston, T., & Cepeda (Wiseheart), N. J. (2013). Laptop multitasking hinders classroom learning for both users and nearby peers. *Computers & Education*, 62, 24–31.

Chapter 14

Table 14.1: Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin*, 138, 353–387.

Appendix C: Statistical Tables

Table C.1: Adapted from Fisher and Yates, *Statistical Tables for Biological, Agricultural, and Medical Research* (1963, 6th ed.), London: Longman.

Figure 1.1 Text Alternative (Chapter 1)

[Return to Figure 1.1](#)

A flow chart presented in a list.

Generate idea.

- Consult past research.
 - State hypothesis.
- Design study.
- Obtain ethics approval.
 - Collect data.
- Analyze data.
- Conduct new study to replicate or extend results (optional).
 - Write manuscript.
- Submit manuscript to peer-reviewed journal.
 - State hypothesis (repeat).

[Return to Figure 1.1](#)

Figure 2.1 Text Alternative (Chapter 2)

[Return to Figure 2.1](#)

A list describing major sections and their purposes.

Tend to start broadly (with a statement of the topic) and narrow toward study method.

- Abstract, a brief summary of the article).
- Introduction, an outline the problem, tie to past research, point to question and method.
- Methodology, detailed description of study design.
- Results, objective report of study results.

Tends to recap results and then provide more general information.

- Discussion, interpretation of study results.
- References, list of all works cited.

[Return to Figure 2.1](#)

Figure 3.1 Text Alternative (Chapter 3)

[Return to Figure 3.1](#)

A flow chart presented in a list.

- Assess potential benefits to participants, science, and society.
 - Assess potential risks to participants.
 - Assess whether and potential benefits outweigh the risks.
 - If yes, carry out the research.
 - If no, study must be modified.

[Return to Figure 3.1](#)

Figure 3.3 Text Alternative (Chapter 3)

[Return to Figure 3.3](#)

A flow chart presented in a list.

Did I write the words?

- Yes.
 - Did I think of the idea?
- Yes.
- No need for citation.
- No.
- Provide citation.
- No.
 - "quote words" and provide citation.

[Return to Figure 3.3](#)

Figure 4.1 Text Alternative (Chapter 4)

[Return to Figure 4.1](#)

Four line graphs.

Graph A, titled Positive Linear Relationship. The x-axis, labelled Narcissism, ranges from low to high. The y-axis, labelled Frequency of Facebook Use, ranges from low to high. The line begins in the lower left and ends in the upper right.

Graph B, titled Negative Linear Relationship. The x-axis, labelled Depressive Symptoms, ranges from low to high. The y-axis, labelled Quality of Advisor Relationship, ranges from low to high. The line begins in the upper left and ends in the lower right.

Graph C, titled Curvilinear Relationship. The x-axis, labelled Years of Teaching, ranges from low to high. The y-axis, labelled Efficacy Beliefs, ranges from low to high. A curve begins in the lower left, peaks just above the center of the graph, and ends in the lower right.

Graph D, titled No Relationship. The x-axis, labelled Automobile Accidents, ranges from low to high. The y-axis, labelled Full Moon Phase, ranges from low to high. A horizontal line begins near the middle of the y-axis and crosses the graph/

[Return to Figure 4.1](#)

Figure 4.2 Text Alternative (Chapter 4)

[Return to Figure 4.2](#)

A line graph. The x-axis, labelled Variable A, ranges from low to high. The y-axis, labelled Variable B, ranges from low to high. The line begins in the lower left, curves up and right, and ends in the upper right.

[Return to Figure 4.2](#)

Figure 4.3 Text Alternative (Chapter 4)

[Return to Figure 4.3](#)

A flow chart presented in a list.

- Inattention.
 - (positive) Behaviour problems.
 - (positive) Peer problems.
 - (negative) Prosocial concern.
 - (negative) Peer problems.

[Return to Figure 4.3](#)

Figure 4.4 Text Alternative (Chapter 4)

[Return to Figure 4.4](#)

Three flow charts. Exercise causes increased happiness. Happiness causes increased exercise. A third variable such as income is associated with both variables, creating an apparent relationship between exercise and happiness. High levels of income result in more exercise; high income also leads to increased happiness.

[Return to Figure 4.4](#)

Figure 4.5 Text Alternative (Chapter 4)

[Return to Figure 4.5](#)

A bar graph. The x-axis, labelled Exposure to Laptop, has two bars: classmates multitasking on laptops and classmates not using laptops. The y-axis, labelled Percent Correct, ranges from 0 to 100 in increments of 10. Classmates multitasking on laptops is at 53%, Classmates not using laptops is at 70%. All data is approximate.

[Return to Figure 4.5](#)

Figure 5.1 Text Alternative (Chapter 5)

[Return to Figure 5.1](#)

A line graph. The x-axis, labelled Score on Test, ranges from 85 to 115 in increments of 15. The y-axis is labelled Number of Scores Obtained. Two normal curves are plotted. The first starts at 85 on the x-axis, peaks at 100 just below halfway up the y-axis, and ends at 115. The second starts at 97, peaks at 100 to the top of the graph, and ends at 103. All data is approximate.

[Return to Figure 5.1](#)

Figure 5.2 Text Alternative (Chapter 5)

[Return to Figure 5.2](#)

A chart consisting of labeled boxes, converted into a list.

Reliability: A reliable measure is consistent. Reliability coefficient: correlation coefficient ranging from 0.00 to 1.00.

Test-Retest Reliability: How consistent is the measure across time?

Take measure two times.

Correlation of score at time one with score at time two; scores should be similar.

Internal Consistency Reliability: How consistent is the measure across items intended to measure the same concept?

Cronbach's alpha: based on correlation of each item on test with every other item.

Interrater Reliability: How consistent is the measure when different people are rating?

Extent to which raters agree in their observations (e.g., using Cohen's kappa).

[Return to Figure 5.2](#)

Figure 5.3 Text Alternative (Chapter 5)

[Return to Figure 5.3](#)

A chart consisting of labeled boxes, converted into a list.

Methods for Building an Argument for Construct Validity.

Does the content of measure reflect the theoretical meaning of the construct?

Face Validity: The content of the measure appears to reflect the construct being measured.

Content Validity: The content of the measure captures all the necessary aspects of the construct and nothing more.

How does this measure relate to other measures and behaviours?

Predictive Validity: Scores on the measure predict behaviour on a criterion measured in the future.

Concurrent Validity: Scores on the measure are related to a criterion (e.g., a behaviour) measured now.

Convergent Validity: Scores on the measure are related to other measures of the same or very similar constructs.

Discriminant Validity: Scores on the measure are not related to other measures that capture theoretically different constructs.

[Return to Figure 5.3](#)

A Flow Chart Planning a Between-Subjects Experiment Graphic Text Alternative (Chapter 8)

[Return to a Flow Chart Planning a Between-Subjects Experiment Graphic](#)

A flow chart presented in a list.

- Participants.
 - Experimental group (independent variable), via assignment to condition.
 - Measure (dependent variable).
 - Control group, (independent variable), via assignment to condition.
 - Measure (dependent variable).

[Return to a Flow Chart Planning a Between-Subjects Experiment Graphic](#)

A Flow Chart Presented Listen then read Graphic Text Alternative (Chapter 8)

[Return to a Flow Chart Presented Listen then Read Graphic](#)

A flow chart presented in a list.

Participants.

- Order 1, via random assignment to condition;
 - Listen (repeated measures independent variable).
 - Memory measure (dependent variable).
 - Read silently (repeated measures independent variable).
 - Memory measure (dependent variable).
 - Order 2, via random assignment to condition.
 - Read silently (repeated measures independent variable).
 - Memory measure (dependent variable).
 - Listen (repeated measures independent variable).
 - Memory measures (dependent variable).

[Return to a Flow Chart Presented Listen then Read Graphic](#)

Figure 10.1 Text Alternative (Chapter 10)

[Return to Figure 10.1](#)

A flow chart, each stage points down to the one following, presented in a list.

- Needs Assessment (Are there problems that need to be addressed in a target population?).
- Program Theory Assessment (How will the problems be addressed? Will the proposed program actually address the needs appropriately?).
- Process Evaluation (Is the program addressing the needs appropriately? Is it being implemented appropriately?)
- Outcome Evaluation (Are the intended outcomes of the program being realized?).
- Efficiency Assessment (Is the cost of the program worth the outcomes?)

[Return to Figure 10.1](#)

A Flow Chart Concerning Independent and Dependent Variables Graphic Text Alternative (Chapter 10)

[Return to a Flow Chart Concerning Independent and Dependent Variables Graphic](#)

Labelled boxes, connected by arrows, are presented in a list:

Participants.

Independent Variable: Sit Next to Stranger.

Dependent Variable: Measure Time until Stranger Leaves.

[Return to a Flow Chart Concerning Independent and Dependent Variables Graphic](#)

A Flow Chart One-Group Pretest-Posttest Design Graphic Text Alternative (Chapter 10)

[Return to a Flow Chart One-Group Pretest-Posttest Design Graphic](#)

Labelled boxes, connected by arrows, are presented in a list:

Participants.

Dependent Variable Pretest: Smoking Measure.

Independent Variable: Training Program.

Dependent Variable Posttest: Smoking Measure.

[Return to a Flow Chart One-Group Pretest-Posttest Design Graphic](#)

Two Flow Charts Non-Equivalent Control Group Design Graphic Text Alternative (Chapter 10)

[Return to Two Flow Charts Non-Equivalent Control Group Design Graphic](#)

Labelled boxes, connected by arrows, are presented in a list:

Participants.

An arrow labelled No Random Assignment.

Independent Variable: Training Program.

Dependent Variable: Smoking Variable.

Participants.

An arrow labelled No Random Assignment.

Independent Variable: No Training Program.

Dependent Variable: Smoking Measure.

[Return to Two Flow Charts Non-Equivalent Control Group Design Graphic](#)

Two Flow Charts Non-Equivalent Control Group Pretest-Posttest Design Graphic Text Alternative (Chapter 10)

[Return to Two Flow Charts Non-Equivalent Control Group Pretest-Posttest Design Graphic](#)

Labelled boxes, connected by arrows, are presented in a list:

Participants.

An arrow labelled No Random Assignment.

Dependent Variable Pretest: Measure.

Independent Variable: Treatment.

Dependent Variable Posttest: Measure.

Participants.

An arrow labelled No Random Assignment.

Dependent Variable Pretest: Measure.

Independent Variable: No Treatment Control.

Dependent Variable Posttest: Measure.

[Return to Two Flow Charts Non-Equivalent Control Group Pretest-Posttest Design Graphic](#)

Figure 10.2 Text Alternative (Chapter 10)

[Return to Figure 10.2](#)

A line graph. The x-axis, labelled year, ranges from 1990 to 1998. A dashed line labeled Treatment extends up from 1997. The y-axis, labelled drivers killed per year, ranges from 0 to 600 in increments of 200. A line begins at (1990, 750), and ends at (1998, 500). All data is approximate.

[Return to Figure 10.2](#)

Figure 10.3 Text Alternative (Chapter 10)

[Return to Figure 10.3](#)

A line graph. The x-axis, labelled year, ranges from 1990 to 1998. A dashed line labelled Treatment extends up from 1997. The y-axis, labelled drivers killed per year, ranges from 0 to 600 in increments of 200. Two lines are plotted: Ontario and two comparable provinces. The Ontario line begins at (1990, 750), and ends at (1998, 500). The comparable provinces line, nearly a straight line, begins at (1990, 185), and ends at (1998, 180). All data is approximate.

[Return to Figure 10.3](#)

Figure 10.4 Text Alternative (Chapter 10)

[Return to Figure 10.4](#)

A line graph titled Mary, female, 4 years. The x-axis, labelled sessions, ranges from 1 to 22. The y-axis, labelled percentage of compliance to selected teacher requests, ranges from 0 to 100% in increments of 25. The graph is split into three sections by vertical lines; Baseline from 1 to 9, treatment from 10 to 14, and Baseline from 15 to 22. The first Baseline line begins at (1, 30), rises to (4, 100), and is jagged until ending at (9, 55). The Treatment line begins at (10, 55), rises to (11, 100), and stays horizontal to (14, 100). The second Baseline line begins at (15, 75), dips to (16, 55), rises to (18, 100), dips to (19, 55), rises to (21, 100), and ends at (22, 50).

[Return to Figure 10.4](#)

Figure 10.5 Text Alternative (Chapter 10)

[Return to Figure 10.5](#)

Four line graphs. The x-axes, labelled sessions, range from 1 to 28. The y-axes, labelled percentage of compliance to selected teacher requests, range from 0 to 100% in increments of 25. All data is approximate.

Graph titled Kate, female, 4 years. The graph is split into three sections by vertical lines; Baseline from 1 to 3, treatment from 4 to 7, and Baseline from 8 to 22. The first baseline line begins at (1, 50), peaks at (2, 70), and ends at (3, 50). The treatment line begins at (4, 100), dips to (5, 55), and remains jagged ending at (8, 100). The second baseline line begins at (9, 55), peaks at (10, 100), and ends at (13, 100).

Graph titled Ben, male, 4 years. The graph is split into three sections by vertical lines; Baseline from 1 to 4, treatment from 5 to 8, and Baseline from 9 to 22. The first baseline line begins at (1, 40), peaks at (2, 80), and ends at (4, 60). The treatment line begins at (4, 100), falls at (7, 75), and ends at (8, 100). The second baseline line begins at (9, 100) and ends at (10, 100), due to data collection stopping due to child absences.

Graph titled Tyler, male, 3 years. The graph is split into three sections by vertical lines; Baseline from 1 to 6, treatment from 7 to 15, and Baseline from 16 to 22. The first baseline line begins at (1, 20), peaks at (5, 100), and ends at (6, 30). The Treatment line begins at (7, 100), falls at (8, 60), and remains jagged ending at (15, 100). The second baseline line begins at (16, 100), falls at (19, 60), and ends at (20, 85).

Graph titled Jake, male, 4 years. The graph is split into three sections by vertical lines; Baseline from 1 to 8, treatment from 9 to 18, and Baseline from 19 to 22.

The first baseline line begins at (1, 20), peaks at (2, 100), dips to (4, 50), and ends at (8, 50). The treatment line begins at (9, 35), peaks at (11, 100), dips to (13, 70), and ends after stagnating from 15 to 18 at 100. The second baseline line begins at (19, 75), is jagged peaking at (20, 100), falling to the lowest point at (25, 30), rising to (26, 100), and ending at (28, 100).

[Return to Figure 10.5](#)

Figure 11.1 Text Alternative (Chapter 11)

[Return to Figure 11.1](#)

A double line graph. The vertical axis is labelled Performance Level (Percent Correct), and ranges from 0 to 100 in increments of 10 (increments 10 through 40 are not displayed). The horizontal axis is labelled Amount of Reward Promised, with the following marked, left to right: No reward, \$10.00, \$20.00, \$30.00, and \$40.00. All data are approximate.

A point is plotted at 50 percent, No reward. A second point is plotted at 100 percent, \$40.00. A diagonal line connects these two points.

A curved, dashed line connects the following points: 50 percent, No reward; 80 percent, \$10.00; 93 percent, \$20.00; 97 percent, \$30.00; and 100 percent, \$40.00.

[Return to Figure 11.1](#)

Figure 11.2 Text Alternative (Chapter 11)

[Return to Figure 11.2](#)

A line graph. The vertical axis is labelled Dependent Variable, and ranges from Low to High. The horizontal axis is labelled Independent Variable, and is marked at Level 1, Level 2, and Level 3. Three points are plotted and connected by a line: Low, Level 1; High, Level 2; and Low, Level 3.

[Return to Figure 11.2](#)

Figure 11.3 Text Alternative (Chapter 11)

[Return to Figure 11.3](#)

A bar graph and a line graph. The x-axes, labelled Confederate's Food Selection Condition, has increments: 2 candies and 30 candies. The y-axes, labelled Average Number of Candies Eaten, ranges from 0 to 12 in increments of 4. All data is approximate. The bar graph has pairs of bars: thin and obese. At 2 candies: Thin is at 3, obese is at 4. At 30 candies: Thin is 10, obese is 6. In the line graph, the data is identical. The Thin line begins at (2 candies, 3) and ends at (30 candies, 10); Obese begins at (2 candies, 4) and ends at (30 candies, 6).

[Return to Figure 11.3](#)

Figure 11.4 Text Alternative (Chapter 11)

[Return to Figure 11.4](#)

8 outcomes, each containing a line graph and a square divided in four quarters. The squares are presented as tables. The x-axes of the line graphs are labelled A, and range from 1 to 2; the y-axes, labelled Dependent Variable, range from 1 to 9 in increments of 4.

1. The line begins at (1, 5), labelled B subscript 1, and ends at (2, 5), labelled B subscript 2.

Table.

	B 1	B 2
A 1	5	5
A 2	5	5
	5	5

Main effect of A: No. Main effect of B: No. Interaction between A times B: No.

2. The line begins at (1, 1) B subscript 2, and ends at (2, 9) B subscript 1.

Table.

	B 1	B 2
A 1	1	1
A 2	9	9
	5	5

Main effect of A: Yes. Main effect of B: No. Interaction between A times B: No.

3. The line B subscript 2 begins at (1, 1) and ends at (2, 1). The line B subscript 1 begins at (1, 9) and ends at (2, 9).

Table.

	B 1	B 2
A 1	9	1
A 2	9	1
	9	1

Main effect of A: No. Main effect of B: Yes. Interaction between A times B: No.

4. The line B subscript 2 begins at (1, 1) and ends at (2, 5). B subscript 1 begins at (1, 5) and ends at (2, 9).

Table.

	B 1	B 2
A 1	5	1
A 2	9	5
	7	3

Main effect of A: Yes. Main effect of B: Yes. Interaction between A times B: No.

5. The line B subscript 2 begins at (1, 1) and ends at (2, 1). B subscript 1 begins at (1, 1) and ends at (9, 2).

Table.

	B 1	B 2
A 1	1	1
	1	1

A	2	9	1	5
	5	1		

Main effect of A: Yes. Main effect of B: Yes. Interaction between A times B: Yes.

6. The line B subscript 1 begins at (1, 1) and ends at (2, 9). B subscript 2 begins at (1, 5) and ends at (2, 5).

Table.

	B	1	B	2
A	1	1	5	3
A	2	9	5	7
	5	5		

Main effect of A: Yes. Main effect of B: No. Interaction between A times B: Yes.

7. The line B subscript 2 begins at (1, 5) and ends at (2, 1). B subscript 1 begins at (1, 5) and ends at (2, 9).

Table.

	B	1	B	2
A	1	5	5	5
A	2	9	1	5
	7	3		

Main effect of A: No. Main effect of B: Yes. Interaction between A times B: Yes.

8. The line B subscript 1 begins at (1, 1) and ends at (2, 9). B subscript 2 begins at (1, 9) and ends at (2, 1).

Table.

	B 1	B 2
A 1	1	9
A 2	9	1
	5	5

Main effect of A: No. Main effect of B: No. Interaction between A times B: Yes.

[Return to Figure 11.4](#)

Figure 11.5 Text Alternative (Chapter 11)

[Return to Figure 11.5](#)

A bar graph. The x-axis has pairs of bars, Truth and Lie, for non-psychopathic and psychopathic. The y-axis, labelled Average Number of Head Movements per 100 seconds, ranges from 0 to 12 in increments of 4. At non-psychopathic: Truth is at 7, lie is at 9. At psychopathic: truth is at 6, lies at 11.

[Return to Figure 11.5](#)

Figure 11.6 Text Alternative (Chapter 11)

[Return to Figure 11.6](#)

Three factorial design tables labelled I: Independent Groups Design; II: Repeated Measures Design; III: Combination of Independent Groups and Repeated Measures Designs. I: Independent Groups Design:

B: 1	B: 1	B: 2	B: 2
A: 1 P subscript 1	P subscript 6	P subscript 11	P subscript 16
A: 1 P subscript 2	P subscript 7	P subscript 12	P subscript 17
A: 1 P subscript 3	P subscript 8	P subscript 13	P subscript 18
A: 1 P subscript 4	P subscript 9	P subscript 14	P subscript 19
A: 1 P subscript 5	P subscript 10	P subscript 15	P subscript 20
A: 2 P subscript 21	P subscript 26	P subscript 31	P subscript 36
A: 2 P subscript 22	P subscript 27	P subscript 32	P subscript 37
A: 2 P subscript 23	P subscript 28	P subscript 33	P subscript 38
A: 2 P subscript 24	P subscript 29	P subscript 34	P subscript 39
A: 2 P subscript 25	P subscript 30	P subscript 35	P subscript 40

II: Repeated Measures Design:

B: 1	B: 1	B: 2	B: 2
A: 1 P subscript 1	P subscript 6	P subscript 1 P subscript 6	P subscript 6
A: 1 P subscript 2	P subscript 7	P subscript 2 P subscript 7	P subscript 7
A: 1 P subscript 3	P subscript 8	P subscript 3 P subscript 8	P subscript 8
A: 1 P subscript 4	P subscript 9	P subscript 4 P subscript 9	P subscript 9
A: 1 P subscript 5	P subscript 10	P subscript 5 P subscript 10	P subscript 10
A: 2 P subscript 1	P subscript 6	P subscript 1 P subscript 6	P subscript 6
A: 2 P subscript 2	P subscript 7	P subscript 2 P subscript 7	P subscript 7

A: 2 P subscript 3 P subscript 8 P subscript 3 P subscript 8
A: 2 P subscript 4 P subscript 9 P subscript 4 P subscript 9
A: 2 P subscript 5 P subscript 10 P subscript 5 P subscript 10

III: Combination of Independent Groups and Repeated Measures Designs:

B: 1	B: 1	B: 2	B: 2
A: 1 P subscript 1	P subscript 6	P subscript 1	P subscript 6
A: 1 P subscript 2	P subscript 7	P subscript 2	P subscript 7
A: 1 P subscript 3	P subscript 8	P subscript 3	P subscript 8
A: 1 P subscript 4	P subscript 9	P subscript 4	P subscript 9
A: 1 P subscript 5	P subscript 10	P subscript 5	P subscript 10
A: 2 P subscript 11	P subscript 16	P subscript 11	P subscript 16
A: 2 P subscript 12	P subscript 17	P subscript 12	P subscript 17
A: 2 P subscript 13	P subscript 18	P subscript 13	P subscript 18
A: 2 P subscript 14	P subscript 19	P subscript 14	P subscript 19
A: 2 P subscript 15	P subscript 20	P subscript 15	P subscript 20

[Return to Figure 11.6](#)

Figure 11.7 Text Alternative (Chapter 11)

[Return to Figure 11.7](#)

A double line graph. The vertical axis is labelled Performance Score, and ranges from 0 to 12 in increments of 1. The horizontal axis is labelled Anxiety Level, and is labelled Low, Moderate, and High. All data are approximate. Points plotted are connected by lines: a solid line labelled Easy Task, and a dashed line labelled Hard Task.

Easy Task:

Performance Score 4, Low.

Performance Score 7, Moderate.

Performance Score 10 , High.

Hard Task:

Performance Score 7, Low.

Performance Score 4, Moderate.

- Performance Score 1, High.

[Return to Figure 11.7](#)

Figure 12.1 Text Alternative (Chapter 12)

[Return to Figure 12.1](#)

A bar graph. The vertical axis is labeled Frequency (percent), and ranges from 0 to 100 in increments of 10. The horizontal axis is labeled Response Options to Identify the Most Serious Problem for the World, and is marked, left to right: People living in poverty and need; Discrimination against girls and women; Poor sanitation and infectious diseases; Inadequate education; Environmental pollution. All data are approximate. The following bars are plotted:

60 percent; People living in poverty and need.

7 percent; Discrimination against girls and women.

10 percent; Poor sanitation and infectious diseases.

11 percent; Inadequate education.

13 percent; Environmental pollution.

[Return to Figure 12.1](#)

Figure 12.2 Text Alternative (Chapter 12)

[Return to Figure 12.2](#)

A bar graph. The x-axis, labelled Age for Completion of Education, ranges from 0 to 35 in increments of 5. The y-axis, labelled Frequency, ranges from 0 to 20,000 in increments of 5,000. All data is approximate. The data remains below 5,000 until 15 on the x-axis, when the data reaches 9,000. At 18 years, the data peaks at 19,000. From 20 to 35 years the data declines from 8,000 down.

[Return to Figure 12.2](#)

Figure 12.3 Text Alternative (Chapter 12)

[Return to Figure 12.3](#)

A bar graph, titled Mean and Standard Deviations, with a normal curve following the bars. The x-axis ranges from 10 to 30 in increments of 5, and is divided by 5 vertical lines. The vertical line at 14 is labelled negative 2 S D; at 17, negative 1 S D; 19 and 20; at 23, positive 1 S D; at 26, positive 2 S D. Below the x-axis at 15 is labelled 14%; below 18, 34%; below 21, 34%; and below 24, 14%. The curve starts at (10, 0), peaks at (20, 1,200), and ends at (30, 0).

[Return to Figure 12.3](#)

Figure 12.4 Text Alternative (Chapter 12)

[Return to Figure 12.4](#)

A triple line graph. The vertical axis is labeled Frequency (percent), and ranges from 0 to 35 in increments of 5. The horizontal axis is labeled Age of Respondent, and is marked, from left to right: 10-19; 20-29; 30-39; 40-49; 50-59; 60-69; 70-79; 80-89; 90-99. All data are approximate. Three sets of points are plotted and connected by lines: World, Australia, and Mexico. All data begin at the origin. The following lines are plotted:

World:

5 percent; 10-19.

25 percent; 20-29.

21 percent; 30-39.

19 percent; 40-49.

15 percent; 50-59.

10 percent; 60-69.

6 percent; 70-79.

2 percent; 80-89.

0 percent; 90-99.

Australia:

4 percent; 10-19.

18 percent; 20-29.

19 percent; 30-39.

20 percent; 40-49.

17 percent; 50-59.

15 percent; 60-69.

8 percent; 70-79.

4 percent; 80-89.

0.5 percent; 90-99.

Mexico:

7 percent; 10-19.

32 percent; 20-29.

22 percent; 30-39.

19 percent; 40-49.

10 percent; 50-59.

5 percent; 60-69.

4 percent; 70-79.

0.5 percent; 80-89.

0 percent; 90-99.

[Return to Figure 12.4](#)

Figure 12.5 Text Alternative (Chapter 12)

[Return to Figure 12.5](#)

Two bar graphs: the left graph has a vertical axis labeled Mean Life Satisfaction (1 equals complete dissatisfied, 10 equals completely satisfied), and ranges from 1 to 10 in increments of 1. The horizontal axis is labeled Respondent Nationality, and is marked Australian and Mexican. All data are approximate. The following bars are plotted:

Australian: 7.

Mexican: 8.5.

The bar graph to the right has a vertical axis labeled Mean Life Satisfaction, and ranges from 7 to 9 in increments of 1. The horizontal axis is labeled Respondent Nationality, and is marked Australian and Mexican. All data are approximate. The following bars are plotted:

Australian: 7.2.

Mexican: 8.5.

[Return to Figure 12.5](#)

Figure 12.6 Text Alternative (Chapter 12)

[Return to Figure 12.6](#)

Two scatterplots labelled: Positive Relationship and Negative Relationship. Both graphs have a vertical axis labeled Variable y, that ranges from 0 to 5 in increments of 1. Both graphs have a horizontal axis labeled Variable x, that ranges from 0 to 5 in increments of 1.

Positive Relationship: Nine points are plotted, beginning at 1 Variable y, 1 Variable x. Each point is plotted up and right in a stepping pattern, ending at 5 Variable y, 5 Variable x.

Negative Relationship: Nine points are plotted, beginning at 5 Variable y, 1 Variable x. Each point is plotted down and right in a stepping pattern, ending at 1 Variable y, 5 Variable x.

[Return to Figure 12.6](#)

Figure 12.7 Text Alternative (Chapter 12)

[Return to Figure 12.7](#)

Four scatterplots. The Positive Relationship (positive .65) plot, Negative Relationship (negative .77) plot, and No Relationship (0.00), share x- and y-axes. The x-axes, labelled Variable x, and the y-axes, labelled Variable y, range from 0 to 5. The fourth is titled Plot Data from Table 12.3 (positive .49). The x-axis, labelled Subjective Physical Health, ranges from 0 to 4. The y-axis, labelled Life Satisfaction, ranges from 0 to 10. No data is plotted. In Positive Relationship, the data is clustered in a swath from (1, 1) to (5, 5). In Negative Relationship, the data is clustered in a swath from (1, 5) to (5, 1). In No Relationship, the data scattered across the graph, with high points of (2, 4) and (4, 4), and low at (3, 1).

[Return to Figure 12.7](#)

Figure 12.8 Text Alternative (Chapter 12)

[Return to Figure 12.8](#)

A scatterplot. The vertical axis is labeled Variable y, and ranges from 0 to 14 in increments of 1, values 1 through 9 are not present. The horizontal axis is labeled Variable x, and ranges from 0 to 6 in increments of 1. All data are approximate. The following points are plotted:

10 Variable y; 1 Variable x.

11 Variable y; 1.5 Variable x.

12 Variable y; 2 Variable x.

13 Variable y; 2.5 Variable x.

14 Variable y; 3 Variable x.

14 Variable y; 4 Variable x.

13 Variable y; 4.5 Variable x.

12 Variable y; 5 Variable x.

11 Variable y; 5.5 Variable x.

10 Variable y; 6 Variable x.

[Return to Figure 12.8](#)

Figure 12.9 Text Alternative (Chapter 12)

[Return to Figure 12.9](#)

Two flow charts titled Partial Correlation equals .29 and Partial Correlation equals .37.

In Partial correlation equals .29, Life satisfaction and Freedom of choice are connected by .38, Freedom of choice and Satisfaction with income are connected by .28, and Satisfaction with income and life satisfaction are connected by .48.

In Partial Correlation equals .37, Life satisfaction and Freedom of choice are connected by .38, Freedom of choice and Age are connected by negative .04, and Age and Life satisfaction are connected by negative .03.

[Return to Figure 12.9](#)

Figure 13.1 Text Alternative (Chapter 13)

[Return to Figure 13.1](#)

A normal curve titled Critical Value for Two-Tailed Test with a .05 Significance Level. The x-axis, labelled t, ranges from negative 4 to positive 4. The y-axis, labelled probability, ranges from 0 to .4 in increments of .1. All data is approximate. The curve begins at (negative 3.4, 0.25), peaks at (0, .4), and ends at (positive 3, .025). The area below the curve between negative 3.4 and negative 2.101, and the area between positive 2.101 and positive 3.4 are labelled .025. The remainder is the area below the curve is labelled .95.

[Return to Figure 13.1](#)

Figure 13.2 Text Alternative (Chapter 13)

[Return to Figure 13.2](#)

A graphic presented in a table.

Decision: Reject the Null Hypothesis	Population: Null Hypothesis is True Type 1 Error (alpha)	Population: Null Hypothesis is False Correct Decision (1 minus beta)
Decision: Retain the Null Hypothesis	Correct Decision (1 minus alpha)	Type 2 Error (beta)

[Return to Figure 13.2](#)

Figure 13.3 Text Alternative (Chapter 13)

[Return to Figure 13.3](#)

A graphic presented in a table.

	True State: Null is True (Innocent)	True State: Null is False (Guilty)
Decision: Reject Null (Find Guilty)	Type 1 Error	Correct Decision
Decision: Retain Null (Find Innocent)	Correct Decision	Type 2 Error

[Return to Figure 13.3](#)

Figure 13.4 Text Alternative (Chapter 13)

[Return to Figure 13.4](#)

A graphic presented in a table.

	True State: Null is True (No Operation Needed)	True State: Null is False (Operation is Needed)
Decision: Reject Null (Operate on Patient)	Type 1 Error	Correct Decision
Decision: Retain Null (Don't Operate)	Correct Decision	Type 2 Error

[Return to Figure 13.4](#)

Figure 14.1 Text Alternative (Chapter 14)

[Return to Figure 14.1](#)

Four line graphs, each x-axis, labelled crowding, ranges from low to high, and each y-axis labelled aggression, ranges from low to high. Graph A, a line begins at (low, low) and ends at (high, high), the area above the line is labelled Rural, below is labelled Urban. Graph B, a line labelled Urban begins at (low, low) and ends at (high, middle); the Rural line is parallel and above the Urban line ending at (high, high). Graph C, the Urban line begins at (low, low) and ends at (high, low); the Rural line begins at (low, low) and ends at (high, high). Graph D, the Rural line begins at (low, low) and ends at (high, high); the Urban line begins at (low, high) and ends at (high, low).

[Return to Figure 14.1](#)

Figure 14.2 Text Alternative (Chapter 14)

[Return to Figure 14.2](#)

A flow chart, each box flows down to the next. Presented in a list.

- Generate Idea, 1 and 2.
- Consult Past Research, 2.
- State Hypothesis, 1 and 2.
- Design Study, 4, 5, 6, 7, 8, 9, 10 and 11.
- Obtain Ethics Approval, 3 and 9.
- Collect Data, 9.
- Analyse Data , 12, 13, and B.
- Conduct New Study to Replicate or Extend Results, 14 (points back to State Hypothesis, 1 and 2).
- Write Manuscript, 2 and A.
- Submit Manuscript to Peer-Reviewed Journal, A.

[Return to Figure 14.2](#)

Figure A.1 Text Alternative (Appendix A)

[Return to Figure A.1](#)

A diagram of a poster layout. The poster measures 6 feet wide by 4 feet tall. At the top centre is a box labelled Title, Authors and Affiliation. Below are eight labelled boxes, two rows of four boxes. Top row, left to right: Abstract, Method, Figure or Table, and Figure or Table. Bottom row, left to right: Introduction, Results, Figure or Table, and Conclusions.

[Return to Figure A.1](#)

Figure 1 Text Alternative (Appendix A)

[Return to Figure 1](#)

A double bar graph. The vertical axis is labelled Negative emotion, and ranges from 0 to 12 in increments of 2. The horizontal axis is labelled Portion condition, and is marked, left to right: Smaller slice, Same slice, and Larger slice; two bars each: Unrestrained and Restrained. All data are approximate.

Smaller slice:

Unrestrained: 9.

Restrained: 11.

Same slice:

Unrestrained: 10.

Restrained: 10.5.

Larger slice:

Unrestrained: 10.5.

Restrained: 10.

[Return to Figure 1](#)

Figure D.1 Text Alternative (Appendix D)

[Return to Figure D.1](#)

Two tables.

	Sex of Subject	Hand Dominance: Right	Hand Dominance: Left	Hand Dominance: Ambidextrous	Row Totals
Male	O subscript 1 equals 15; E subscript 1 equals 25	O subscript 2 equals 30; E subscript 2 equals 20	O subscript 3 equals 5; E subscript 3 equals 5	50	
	O subscript 4 equals 35; E subscript 4 equals 25	O subscript 5 equals 10; E subscript 5 equals 20	O subscript 6 equals 5; E subscript 6 equals 5.		50
Female					

[Return to Figure D.1](#)

Table of Contents

1. [Table of Contents and Preface](#)
 1. [Cover Page](#)
 2. [Title Page](#)
 3. [Copyright Information](#)
 4. [Dedication](#)
 5. [About the Authors](#)
 6. [Brief Contents](#)
 7. [Contents](#)
 8. [Preface](#)
 1. [Features of This Canadian Edition](#)
 2. [Organization & Other changes](#)
 3. [Award-Winning Technology](#)
 4. [Acknowledgments](#)
 9. [Connect](#)
2. [Chapter 1: Scientific Understanding of Behaviour](#)
 1. [Chapter 1 Introduction](#)
 2. [Why Study Research Methods?](#)
 3. [Methods of Acquiring Knowledge](#)
 1. [Intuition](#)
 2. [Authority](#)
 3. [LO2 The Scientific Method: Be Skeptical, Seek Empirical Data](#)
 4. [Science as a Way to Ask and Answer Questions](#)
 4. [Goals of Scientific Research in Psychology](#)
 1. [Describing Behaviour](#)
 2. [Predicting Behaviour](#)
 3. [Determining the Causes of Behaviour](#)
 4. [Explaining Behaviour](#)
 5. [Basic and Applied Research](#)
 1. [Basic Research](#)
 2. [Applied Research](#)
 3. [Integrating Basic and Applied Research](#)
 6. [Study Terms](#)

7. [Review Questions](#)
8. [Deepen Your Understanding](#)
3. [Chapter 2: Where to Start](#)
 1. [Chapter 2 Introduction](#)
 2. [Where Do Research Ideas Come From?](#)
 1. [Questioning Common Assumptions](#)
 2. [Observation of the World around Us](#)
 3. [Practical Problems](#)
 4. [LO2 Theories](#)
 5. [Past Research](#)
 3. [How Do We Find Out What is Already Known?](#)
 1. [What to Expect in a Research Article](#)
 2. [Other Types of Articles: Literature Reviews and Meta-analyses](#)
 3. [Reading Articles](#)
 4. [LO4 Where Are These Articles Published? An Orientation to Journals and Finding Articles](#)
 4. [Developing Hypotheses and Predictions](#)
 5. [Study Terms](#)
 6. [Review Questions](#)
 7. [Deepen Your Understanding](#)
4. [Chapter 3: Ethical Research](#)
 1. [Chapter 3 Introduction](#)
 2. [Were Milgram's Obedience Experiments Ethical?](#)
 3. [Ethical Research in Canada](#)
 1. [The Tri-Council and Its Policy Statement](#)
 2. [Historical, Legal, and International Context](#)
 3. [LO1 Core Principles Guiding Research with Human Participants](#)
 4. [Designing Research to Uphold the Core Principles](#)
 1. [LO2 Promote Concern for Welfare by Minimizing Risks and Maximizing Benefits](#)
 2. [LO3 Promote Respect for Persons through Informed Consent](#)
 3. [Promote Justice by Involving People Equitably in Research](#)
 4. [Evaluating the Ethics of Research with Human Participants](#)
 5. [Monitoring Ethical Standards at Each Institution](#)

1. [LO6 Exempt Research](#)
 2. [Minimal Risk Research](#)
 3. [Greater Than Minimal Risk Research](#)
 6. [Ethics and Animal Research](#)
 7. [Professional Ethics in Academic Life](#)
 1. [Ethics Codes of the APA and CPA](#)
 2. [LO8 Scientific Misconduct and Publication Ethics](#)
 3. [Plagiarism and the Integrity of Academic Communication](#)
 8. [Study Terms](#)
 9. [Review Questions](#)
 10. [Deepen Your Understanding](#)
5. [Chapter 4: Research Design Fundamentals](#)
1. [Chapter 4 Introduction](#)
 2. [Introduction to Basic Research Design](#)
 1. [Variables](#)
 2. [LO1 Two Basic Research Designs](#)
 3. [LO2 Operationally Defining Variables: Turning Hypotheses into Predictions](#)
 3. [Non-experimental Method](#)
 1. [LO4 Relationships between Variables](#)
 2. [Interpreting the Results of Non-experimental designs](#)
 4. [Experimental Method](#)
 1. [LO6 Designing Experiments That Allow for Causal Inferences](#)
5. [Choosing a Method: Advantages of Multiple Methods](#)
1. [Artificiality of Experiments](#)
 2. [Ethical and Practical Considerations](#)
 3. [Describing Behaviour](#)
 4. [Predicting Future Behaviour](#)
 5. [Advantages of Multiple Methods](#)
6. [Study Terms](#)
 7. [Review Questions](#)
 8. [Deepen Your Understanding](#)
6. [Chapter 5: Measurement](#)
1. [Chapter 5 Introduction](#)
 2. [Self-Report Measures](#)
 3. [Reliability](#)

1. [Test-Retest Reliability](#)
2. [Internal Consistency Reliability](#)
3. [Inter-rater Reliability](#)
4. [LO2 Reliability and Accuracy of Measures](#)
4. [Validity of Measures](#)
 1. [Indicators of Construct Validity](#)
5. [Reactivity of Measures](#)
6. [Variables and Measurement Scales](#)
 1. [Nominal Scales](#)
 2. [Ordinal Scales](#)
 3. [Interval Scales](#)
 4. [Ratio Scales](#)
 5. [The Importance of the Measurement Scales](#)
7. [Study Terms](#)
8. [Review Questions](#)
9. [Deepen Your Understanding](#)
7. [Chapter 6: Observational Methods](#)
 1. [Chapter 6 Introduction](#)
 2. [Quantitative and Qualitative Approaches](#)
 3. [Naturalistic Observation](#)
 1. [Issues in Naturalistic Observation](#)
 4. [Systematic Observation](#)
 1. [Coding Schemes](#)
 2. [Issues in Systematic Observation](#)
 5. [Case Studies](#)
 6. [Archival Research](#)
 1. [Census Data or Statistical Records](#)
 2. [Survey Archives](#)
 3. [Written Records and Mass Media](#)
 4. [Working with Archival Data: Content Analysis and Interpretation](#)
 7. [Study Terms](#)
 8. [Review Questions](#)
 9. [Deepen Your Understanding](#)
8. [Chapter 7: Survey Research: Asking People about Themselves](#)
 1. [Chapter 7 Introduction](#)
 2. [Why Conduct Surveys?](#)

1. [Response Bias in Survey Research](#)
3. [Constructing Good Questions](#)
 1. [Defining the Research Objectives](#)
 2. [Question Wording](#)
4. [Responses to Questions: What Kind of Data Are You Seeking?](#)
 1. [Closed- versus Open-Ended Questions](#)
 2. [Rating Scales for Closed-Ended Questions](#)
5. [Finalizing the Questionnaire](#)
 1. [Formatting the Questionnaire](#)
 2. [Refining Questions](#)
6. [Administering Surveys](#)
 1. [Questionnaires](#)
 2. [Interviews](#)
7. [Interpreting Survey Results: Consider the Sample](#)
 1. [Population and Samples](#)
 2. [LO5 For More Precise Estimates, Use a Larger Sample](#)
 3. [LO6 To Describe a Specific Population, Sample Thoroughly](#)
8. [Sampling Techniques](#)
 1. [Probability Sampling](#)
 2. [Non-probability Sampling](#)
 3. [Reasons for Using Convenience Samples](#)
9. [Study Terms](#)
10. [Review Questions](#)
11. [Deepen Your Understanding](#)
9. [Chapter 8: Experimental Design](#)
 1. [Chapter 8 Introduction](#)
 2. [Confounding and Internal Validity](#)
 3. [Planning a Basic Experiment](#)
 4. [Between-Subjects Experiments](#)
 1. [LO3 Pretest-Posttest Design](#)
 2. [LO4 Matched Pairs Design](#)
 5. [Within-Subjects Experiments](#)
 1. [LO5 Advantages and Disadvantages of the Within-Subjects Design](#)
 2. [Counterbalancing](#)
 3. [Time Interval between Treatments](#)

4. [Choosing between Between-Subjects and Within-Subjects Designs](#)
 6. [Study Terms](#)
 7. [Review Questions](#)
 8. [Deepen Your Understanding](#)
10. [Chapter 9: Conducting Studies](#)
1. [Chapter 9 Introduction](#)
 2. [Finalizing a Study Design](#)
 1. [LO1 Options for Manipulating the Independent Variable in Experiments](#)
 2. [Additional Considerations when Manipulating the Independent Variable](#)
 3. [LO2 Options for Measuring Variables](#)
 4. [LO3 Additional Considerations when Measuring Variables](#)
 5. [LO4 Setting the Stage](#)
 3. [Advanced Considerations for Ensuring Control](#)
 1. [Controlling for Participant Expectations](#)
 2. [Controlling for Experimenter Expectations](#)
 4. [Seeking Ethics Approval](#)
 1. [Selecting Research Participants](#)
 2. [Planning the Debriefing](#)
 5. [Collecting Data](#)
 1. [Pilot Studies](#)
 2. [Researcher Commitments](#)
 6. [What Comes Next?](#)
 1. [Analyzing and Interpreting Results](#)
 2. [Communicating Research to Others](#)
 7. [Study Terms](#)
 8. [Review Questions](#)
 9. [Deepen Your Understanding](#)
11. [Chapter 10: Research Designs for Special Circumstances](#)
1. [Chapter 10 Introduction](#)
 2. [Program Evaluation](#)
 3. [Quasi-Experimental Designs](#)
 1. [LO2 One-Group Posttest-Only Design](#)
 2. [One-Group Pretest-Posttest Design](#)
 3. [LO3 Threats to Internal Validity](#)

4. [LO4 Non-equivalent Control Group Design](#)
 5. [Non-equivalent Control Group Pretest-Posttest Design](#)
 6. [LO5 Interrupted Time Series Design](#)
 7. [Control Series Design](#)
 8. [Summing up Quasi-Experimental Designs](#)
 4. [Single Case Experimental Designs](#)
 1. [Reversal Designs](#)
 2. [Multiple Baseline Designs](#)
 3. [Replications in Single Case Designs](#)
 5. [Developmental Research Designs](#)
 1. [Longitudinal Method](#)
 2. [Cross-Sectional Method](#)
 3. [Comparing Longitudinal and Cross-Sectional Methods](#)
 4. [Sequential Method](#)
 6. [Study Terms](#)
 7. [Review Questions](#)
 8. [Deepen Your Understanding](#)
12. [Chapter 11: Complex Experimental Designs](#)
1. [Chapter 11 Introduction](#)
 2. [An Independent Variable with More Than Two Levels](#)
 3. [An Experiment with More Than One Independent Variable: Factorial Designs](#)
 1. [LO2 Interpreting Factorial Designs](#)
 2. [Interactions Illuminate Moderator Variables](#)
 3. [LO3 Depicting Possible Outcomes of a \$2 \times 2\$ Factorial Design Using Tables and Graphs](#)
 4. [LO4 Breaking Down Interactions into Simple Main Effects](#)
 4. [Variations on \$2 \times 2\$ Factorial Designs](#)
 1. [LO5 Factorial Designs with Manipulated and Non-manipulated Variables](#)
 2. [LO6 Assignment Procedures and Sample Size](#)
 5. [Increasing the Complexity of Factorial Designs](#)
 1. [Beyond Two Levels per Independent Variable](#)
 2. [Beyond Two Independent Variables](#)
 6. [Study Terms](#)
 7. [Review Questions](#)
 8. [Deepen Your Understanding](#)

13. [Chapter 12: Descriptive Statistics: Describing Variables and the Relations among Them](#)

1. [Chapter 12 Introduction](#)
2. [Revisiting Scales of Measurement](#)
3. [Describing Each Variable](#)
 1. [LO1 Graphing Frequency Distributions](#)
 2. [LO2 Descriptive Statistics](#)
4. [Describing Relationships Involving Nominal Variables](#)
 1. [Comparing Groups of Participants](#)
 2. [Graphing Nominal Data](#)
 3. [LO4 Describing Effect-Size between Two Groups](#)
5. [Describing Relationships among Continuous Variables: Correlating Two Variables](#)
 1. [Interpreting the Pearson r Correlation Coefficient](#)
 2. [Scatterplots](#)
 3. [Important Considerations](#)
 4. [Correlation Coefficients as Effect-Sizes](#)
6. [Describing Relationships among Continuous Variables: Increasing Complexity](#)
 1. [LO6 The Regression Equation](#)
 2. [Multiple Correlation and Multiple Regression](#)
 3. [Integrating Results from Different Analyses](#)
 4. [LO7 Partial Correlation and the Third-Variable Problem](#)
 5. [Advanced Modelling Techniques](#)
7. [Combining Descriptive and Inferential Statistics](#)
8. [Study Terms](#)
9. [Review Questions](#)
10. [Deepen Your Understanding](#)

14. [Chapter 13: Inferential Statistics: Making Inferences about Populations Based on Our Samples](#)

1. [Chapter 13 Introduction](#)
2. [Inferential Statistics: Using Samples to Make Inferences about Populations](#)
 1. [Inferential Statistics: Ruling Out Chance](#)
 2. [Statistical Significance: An Overview](#)
3. [Null and Research Hypotheses](#)
4. [Probability and Sampling Distributions](#)

1. [Probability: The Case of Mind Reading](#)
 2. [LO3 Sampling Distributions](#)
 3. [Sample Size](#)
 4. [How “Unlikely” Is Enough? Choosing a STATISTICAL Significance Level \(Alpha\)](#)
 5. [Example Statistical Tests](#)
 1. [LO4 Thet-Test: Comparing Two Means](#)
 2. [LO5 TheFTest: Used When Comparing Three or More Group Means](#)
 3. [Statistical Significance of a Pearson r Correlation Coefficient](#)
 6. [We Made a Decision about the Null Hypothesis, but We Might Be Wrong! Investigating Type I and Type II Errors](#)
 1. [Correct Decisions](#)
 2. [Type I Errors](#)
 3. [Type II Errors](#)
 4. [The Everyday Context of Type I and Type II Errors](#)
 5. [Type I and Type II Errors in the Published Research Literature](#)
 7. [Interpreting Statistically Non-significant Results](#)
 8. [Choosing a Sample Size: Power Analysis](#)
 9. [Analyzing Data Using Statistics Software](#)
 10. [Selecting the Appropriate Statistical Test](#)
 1. [Research Studying Two Variables](#)
 2. [Research with Multiple Independent or Predictor Variables](#)
 11. [Integrating Descriptive and Inferential Statistics](#)
 1. [Effect-Size](#)
 2. [Confidence Intervals and Statistical Significance](#)
 3. [Conclusion Validity](#)
 12. [The Importance of Replications](#)
 13. [Study Terms](#)
 14. [Review Questions](#)
 15. [Deepen Your Understanding](#)
15. [Chapter 14: Generalizing Results](#)
 1. [Chapter 14 Introduction](#)
 2. [Challenges to Generalizing Results](#)
 1. [LO1 Can Results Generalize to Other Populations?](#)

2. [LO2 Can Results Generalize beyond the Specific Study Situation?](#)
3. [Solutions to Generalizing Results](#)
 1. [LO3 Replicate the Study](#)
 2. [Consider Different Populations](#)
 3. [LO5 Rely on Multiple Studies to Draw Conclusions: Literature Reviews and Meta-analyses](#)
4. [Generalizing Your Knowledge beyond This Book](#)
 1. [Recognize and Use Your New Knowledge](#)
 2. [Stay Connected to Building a Better Psychological Science](#)
 3. [Use Research to Improve Lives](#)
 5. [Study Terms](#)
 6. [Review Questions](#)
 7. [Deepen Your Understanding](#)
16. [Appendix A: Writing Research Reports in APA Style](#)
 1. [Appendix A Introduction](#)
 2. [Writing Style and Word Choice](#)
 1. [Clarity and Intended Audience](#)
 2. [Paraphrase and Cite Past Research](#)
 3. [Active versus Passive Voice](#)
 4. [Avoiding Biased Language](#)
 3. [Writing Each Section of the APA Style Research Report](#)
 1. [Title Page](#)
 2. [Abstract](#)
 3. [Introduction](#)
 4. [Method](#)
 5. [Results](#)
 6. [Discussion](#)
 7. [References](#)
 8. [Appendix](#)
 9. [Author Note](#)
 10. [Footnotes and Endnotes](#)
 11. [Tables](#)
 12. [Figures](#)
 13. [Summary: Order of Pages](#)
 4. [Formatting a Manuscript](#)
 1. [Using Headings](#)

- 2. [Abbreviations](#)
 - 3. [Reporting Numbers and Statistics](#)
 - 4. [APA Style and Student Paper Formats](#)
 - 5. [Citing and Referencing Sources](#)
 - 1. [Citing Sources in the Body of the Report](#)
 - 2. [Reference List Style](#)
 - 3. [Reference Formats for Electronic Sources](#)
 - 6. [Conference Presentations](#)
 - 1. [Paper Presentations](#)
 - 2. [Poster Sessions](#)
 - 7. [Appendix A Sample Paper](#)
17. [Appendix B: Statistical Tests](#)
- 1. [Appendix B Introduction](#)
 - 2. [Descriptive Statistics](#)
 - 1. [Measures of Central Tendency](#)
 - 2. [Measures of Variability](#)
 - 3. [Pearson Product-Moment Correlation Coefficient](#)
 - 3. [Additional Statistical Significance Tests](#)
 - 1. [Significance of Correlation Coefficient r](#)
 - 2. [Chi-Square \(\$\chi^2\$ \)](#)
 - 3. [Analysis of Variance \(F Test\): Overview](#)
 - 4. [Analysis of Variance: One Independent Variable, Between-Subjects Design](#)
 - 5. [Analysis of Variance: Two Independent Variables, Between-Subjects Design](#)
 - 6. [Analysis of Variance: One Independent Variable, Within-subjects Design](#)
 - 7. [Analysis of Variance: Conclusion](#)
 - 4. [Effect Size](#)
 - 1. [Effect Size as Strength of Association](#)
 - 2. [Effect Size as Proportion of Variance Explained](#)
 - 3. [Effect Size as a Standardized Difference between Means](#)
18. [Appendix C: Statistical Tables](#)
- 1. [Appendix C Introduction](#)
19. [Appendix D: How to Conduct a PsycINFO Search](#)
- 1. [Appendix D Introduction](#)
 - 2. [The User Screen](#)

3. [Specifying Search Terms and Finding Results](#)
4. [Combining Search Terms and Narrowing Results](#)
5. [Saving Results](#)
20. [Appendix E: Constructing a Latin Square](#)
 1. [Appendix E Introduction](#)
21. [References](#)
 1. [References](#)
 2. [Sources](#)
22. [Accessibility Content: Text Alternatives for Images](#)
 1. [Figure 1.1 Text Alternative \(Chapter 1\)](#)
 2. [Figure 2.1 Text Alternative \(Chapter 2\)](#)
 3. [Figure 3.1 Text Alternative \(Chapter 3\)](#)
 4. [Figure 3.3 Text Alternative \(Chapter 3\)](#)
 5. [Figure 4.1 Text Alternative \(Chapter 4\)](#)
 6. [Figure 4.2 Text Alternative \(Chapter 4\)](#)
 7. [Figure 4.3 Text Alternative \(Chapter 4\)](#)
 8. [Figure 4.4 Text Alternative \(Chapter 4\)](#)
 9. [Figure 4.5 Text Alternative \(Chapter 4\)](#)
 10. [Figure 5.1 Text Alternative \(Chapter 5\)](#)
 11. [Figure 5.2 Text Alternative \(Chapter 5\)](#)
 12. [Figure 5.3 Text Alternative \(Chapter 5\)](#)
 13. [A Flow Chart Planning a Between-Subjects Experiment Text Alternative \(Chapter 8\)](#)
 14. [A Flow Chart Presented Listen then Read Text Alternative \(Chapter 8\)](#)
 15. [Figure 10.1 Text Alternative \(Chapter 10\)](#)
 16. [A Flow Chart Concerning Independent and Dependent Variables Text Alternative \(Chapter 10\)](#)
 17. [A Flow Chart One-Group Pretest-Posttest Design Text Alternative \(Chapter 10\)](#)
 18. [Two Flow Charts Non-Equivalent Control Group Design Text Alternative \(Chapter 10\)](#)
 19. [Two Flow Charts Non-Equivalent Control Group Pretest-Posttest Design Text Alternative \(Chapter 10\)](#)
 20. [Figure 10.2 Text Alternative \(Chapter 10\)](#)
 21. [Figure 10.3 Text Alternative \(Chapter 10\)](#)
 22. [Figure 10.4 Text Alternative \(Chapter 10\)](#)

23. [Figure 10.5 Text Alternative \(Chapter 10\)](#).
 24. [Figure 11.1 Text Alternative \(Chapter 11\)](#).
 25. [Figure 11.2 Text Alternative \(Chapter 11\)](#).
 26. [Figure 11.3 Text Alternative \(Chapter 11\)](#).
 27. [Figure 11.4 Text Alternative \(Chapter 11\)](#).
 28. [Figure 11.5 Text Alternative \(Chapter 11\)](#).
 29. [Figure 11.6 Text Alternative \(Chapter 11\)](#).
 30. [Figure 11.7 Text Alternative \(Chapter 11\)](#).
 31. [Figure 12.1 Text Alternative \(Chapter 12\)](#).
 32. [Figure 12.2 Text Alternative \(Chapter 12\)](#).
 33. [Figure 12.3 Text Alternative \(Chapter 12\)](#).
 34. [Figure 12.4 Text Alternative \(Chapter 12\)](#).
 35. [Figure 12.5 Text Alternative \(Chapter 12\)](#).
 36. [Figure 12.6 Text Alternative \(Chapter 12\)](#).
 37. [Figure 12.7 Text Alternative \(Chapter 12\)](#).
 38. [Figure 12.8 Text Alternative \(Chapter 12\)](#).
 39. [Figure 12.9 Text Alternative \(Chapter 12\)](#).
 40. [Figure 13.1 Text Alternative \(Chapter 13\)](#).
 41. [Figure 13.2 Text Alternative \(Chapter 13\)](#).
 42. [Figure 13.3 Text Alternative \(Chapter 13\)](#).
 43. [Figure 13.4 Text Alternative \(Chapter 13\)](#).
 44. [Figure 14.1 Text Alternative \(Chapter 14\)](#).
 45. [Figure 14.2 Text Alternative \(Chapter 14\)](#).
 46. [Figure A.1 Text Alternative \(Appendix A\)](#).
 47. [Figure 1 Text Alternative \(Appendix A\)](#).
 48. [Figure D.1 Text Alternative \(Appendix D\)](#).
-
1. [Page i](#)
 2. [Page ii](#)
 3. [Page iii](#)
 4. [Page iv](#)
 5. [Page v](#)
 6. [Page vi](#)
 7. [Page vii](#)
 8. [Page viii](#)
 9. [Page ix](#)
 10. [Page x](#)

11. [Page xi](#)
12. [Page xii](#)
13. [Page xiii](#)
14. [Page xiv](#)
15. [Page xv](#)
16. [Page xvi](#)
17. [Page xvii](#)
18. [Page 1](#)
19. [Page 2](#)
20. [Page 3](#)
21. [Page 4](#)
22. [Page 5](#)
23. [Page 6](#)
24. [Page 7](#)
25. [Page 8](#)
26. [Page 9](#)
27. [Page 10](#)
28. [Page 11](#)
29. [Page 12](#)
30. [Page 13](#)
31. [Page 14](#)
32. [Page 15](#)
33. [Page 16](#)
34. [Page 17](#)
35. [Page 18](#)
36. [Page 19](#)
37. [Page 20](#)
38. [Page 21](#)
39. [Page 22](#)
40. [Page 23](#)
41. [Page 24](#)
42. [Page 25](#)
43. [Page 26](#)
44. [Page 27](#)
45. [Page 28](#)
46. [Page 29](#)
47. [Page 30](#)

48. [Page 31](#)
49. [Page 32](#)
50. [Page 33](#)
51. [Page 34](#)
52. [Page 35](#)
53. [Page 36](#)
54. [Page 37](#)
55. [Page 38](#)
56. [Page 39](#)
57. [Page 40](#)
58. [Page 41](#)
59. [Page 42](#)
60. [Page 43](#)
61. [Page 44](#)
62. [Page 45](#)
63. [Page 46](#)
64. [Page 47](#)
65. [Page 48](#)
66. [Page 49](#)
67. [Page 50](#)
68. [Page 51](#)
69. [Page 52](#)
70. [Page 53](#)
71. [Page 54](#)
72. [Page 55](#)
73. [Page 56](#)
74. [Page 57](#)
75. [Page 58](#)
76. [Page 59](#)
77. [Page 60](#)
78. [Page 61](#)
79. [Page 62](#)
80. [Page 63](#)
81. [Page 64](#)
82. [Page 65](#)
83. [Page 66](#)
84. [Page 67](#)

85. [Page 68](#)
86. [Page 69](#)
87. [Page 70](#)
88. [Page 71](#)
89. [Page 72](#)
90. [Page 73](#)
91. [Page 74](#)
92. [Page 75](#)
93. [Page 76](#)
94. [Page 77](#)
95. [Page 78](#)
96. [Page 79](#)
97. [Page 80](#)
98. [Page 81](#)
99. [Page 82](#)
100. [Page 83](#)
101. [Page 84](#)
102. [Page 85](#)
103. [Page 86](#)
104. [Page 87](#)
105. [Page 88](#)
106. [Page 89](#)
107. [Page 90](#)
108. [Page 91](#)
109. [Page 92](#)
110. [Page 93](#)
111. [Page 94](#)
112. [Page 95](#)
113. [Page 96](#)
114. [Page 97](#)
115. [Page 98](#)
116. [Page 99](#)
117. [Page 100](#)
118. [Page 101](#)
119. [Page 102](#)
120. [Page 103](#)
121. [Page 104](#)

- 122. [Page 105](#)
- 123. [Page 106](#)
- 124. [Page 107](#)
- 125. [Page 108](#)
- 126. [Page 109](#)
- 127. [Page 110](#)
- 128. [Page 111](#)
- 129. [Page 112](#)
- 130. [Page 113](#)
- 131. [Page 114](#)
- 132. [Page 115](#)
- 133. [Page 116](#)
- 134. [Page 117](#)
- 135. [Page 118](#)
- 136. [Page 119](#)
- 137. [Page 120](#)
- 138. [Page 121](#)
- 139. [Page 122](#)
- 140. [Page 123](#)
- 141. [Page 124](#)
- 142. [Page 125](#)
- 143. [Page 126](#)
- 144. [Page 127](#)
- 145. [Page 128](#)
- 146. [Page 129](#)
- 147. [Page 130](#)
- 148. [Page 131](#)
- 149. [Page 132](#)
- 150. [Page 133](#)
- 151. [Page 134](#)
- 152. [Page 135](#)
- 153. [Page 136](#)
- 154. [Page 137](#)
- 155. [Page 138](#)
- 156. [Page 139](#)
- 157. [Page 140](#)
- 158. [Page 141](#)

- 159. [Page 142](#)
- 160. [Page 143](#)
- 161. [Page 144](#)
- 162. [Page 145](#)
- 163. [Page 146](#)
- 164. [Page 147](#)
- 165. [Page 148](#)
- 166. [Page 149](#)
- 167. [Page 150](#)
- 168. [Page 151](#)
- 169. [Page 152](#)
- 170. [Page 153](#)
- 171. [Page 154](#)
- 172. [Page 155](#)
- 173. [Page 156](#)
- 174. [Page 157](#)
- 175. [Page 158](#)
- 176. [Page 159](#)
- 177. [Page 160](#)
- 178. [Page 161](#)
- 179. [Page 162](#)
- 180. [Page 163](#)
- 181. [Page 164](#)
- 182. [Page 165](#)
- 183. [Page 166](#)
- 184. [Page 167](#)
- 185. [Page 168](#)
- 186. [Page 169](#)
- 187. [Page 170](#)
- 188. [Page 171](#)
- 189. [Page 172](#)
- 190. [Page 173](#)
- 191. [Page 174](#)
- 192. [Page 175](#)
- 193. [Page 176](#)
- 194. [Page 177](#)
- 195. [Page 178](#)

196. [Page 179](#)
197. [Page 180](#)
198. [Page 181](#)
199. [Page 182](#)
200. [Page 183](#)
201. [Page 184](#)
202. [Page 185](#)
203. [Page 186](#)
204. [Page 187](#)
205. [Page 188](#)
206. [Page 189](#)
207. [Page 190](#)
208. [Page 191](#)
209. [Page 192](#)
210. [Page 193](#)
211. [Page 194](#)
212. [Page 195](#)
213. [Page 196](#)
214. [Page 197](#)
215. [Page 198](#)
216. [Page 199](#)
217. [Page 200](#)
218. [Page 201](#)
219. [Page 202](#)
220. [Page 203](#)
221. [Page 204](#)
222. [Page 205](#)
223. [Page 206](#)
224. [Page 207](#)
225. [Page 208](#)
226. [Page 209](#)
227. [Page 210](#)
228. [Page 211](#)
229. [Page 212](#)
230. [Page 213](#)
231. [Page 214](#)
232. [Page 215](#)

- 233. [Page 216](#)
- 234. [Page 217](#)
- 235. [Page 218](#)
- 236. [Page 219](#)
- 237. [Page 220](#)
- 238. [Page 221](#)
- 239. [Page 222](#)
- 240. [Page 223](#)
- 241. [Page 224](#)
- 242. [Page 225](#)
- 243. [Page 226](#)
- 244. [Page 227](#)
- 245. [Page 228](#)
- 246. [Page 229](#)
- 247. [Page 230](#)
- 248. [Page 231](#)
- 249. [Page 232](#)
- 250. [Page 233](#)
- 251. [Page 234](#)
- 252. [Page 235](#)
- 253. [Page 236](#)
- 254. [Page 237](#)
- 255. [Page 238](#)
- 256. [Page 239](#)
- 257. [Page 240](#)
- 258. [Page 241](#)
- 259. [Page 242](#)
- 260. [Page 243](#)
- 261. [Page 244](#)
- 262. [Page 245](#)
- 263. [Page 246](#)
- 264. [Page 247](#)
- 265. [Page 248](#)
- 266. [Page 249](#)
- 267. [Page 250](#)
- 268. [Page 251](#)
- 269. [Page 252](#)

- 270. [Page 253](#)
- 271. [Page 254](#)
- 272. [Page 255](#)
- 273. [Page 256](#)
- 274. [Page 257](#)
- 275. [Page 258](#)
- 276. [Page 259](#)
- 277. [Page 260](#)
- 278. [Page 261](#)
- 279. [Page 262](#)
- 280. [Page 263](#)
- 281. [Page 264](#)
- 282. [Page 265](#)
- 283. [Page 266](#)
- 284. [Page 267](#)
- 285. [Page 268](#)
- 286. [Page 269](#)
- 287. [Page 270](#)
- 288. [Page 271](#)
- 289. [Page 272](#)
- 290. [Page 273](#)
- 291. [Page 274](#)
- 292. [Page 275](#)
- 293. [Page 276](#)
- 294. [Page 277](#)
- 295. [Page 278](#)
- 296. [Page 279](#)
- 297. [Page 280](#)
- 298. [Page 281](#)
- 299. [Page 282](#)
- 300. [Page 283](#)
- 301. [Page 284](#)
- 302. [Page 285](#)
- 303. [Page 286](#)
- 304. [Page 287](#)
- 305. [Page 288](#)
- 306. [Page 289](#)

- 307. [Page 290](#)
- 308. [Page 291](#)
- 309. [Page 292](#)
- 310. [Page 293](#)
- 311. [Page 294](#)
- 312. [Page 295](#)
- 313. [Page 296](#)
- 314. [Page 297](#)
- 315. [Page 298](#)
- 316. [Page 299](#)
- 317. [Page 300](#)
- 318. [Page 301](#)
- 319. [Page 302](#)
- 320. [Page 303](#)
- 321. [Page 304](#)
- 322. [Page 305](#)
- 323. [Page 306](#)
- 324. [Page 307](#)
- 325. [Page 308](#)
- 326. [Page 309](#)
- 327. [Page 310](#)
- 328. [Page 311](#)
- 329. [Page 312](#)
- 330. [Page 313](#)
- 331. [Page 314](#)
- 332. [Page 315](#)
- 333. [Page 316](#)
- 334. [Page 317](#)
- 335. [Page 318](#)
- 336. [Page 319](#)
- 337. [Page 320](#)
- 338. [Page 321](#)
- 339. [Page 322](#)
- 340. [Page 323](#)
- 341. [Page 324](#)
- 342. [Page 325](#)
- 343. [Page 326](#)

- 344. [Page 327](#)
- 345. [Page 328](#)
- 346. [Page 329](#)
- 347. [Page 330](#)
- 348. [Page 331](#)
- 349. [Page 332](#)
- 350. [Page 334](#)
- 351. [Page 335](#)
- 352. [Page 336](#)
- 353. [Page 337](#)
- 354. [Page 338](#)
- 355. [Page 339](#)
- 356. [Page 340](#)
- 357. [Page 341](#)
- 358. [Page 342](#)
- 359. [Page 343](#)
- 360. [Page 344](#)
- 361. [Page 345](#)
- 362. [Page 346](#)
- 363. [Page 347](#)
- 364. [Page 348](#)
- 365. [Page 349](#)
- 366. [Page 350](#)
- 367. [Page 351](#)
- 368. [Page 352](#)
- 369. [Page 353](#)
- 370. [Page 354](#)
- 371. [Page 355](#)
- 372. [Page 358](#)
- 373. [Page 359](#)
- 374. [Page 360](#)
- 375. [Page 361](#)
- 376. [Page 362](#)
- 377. [Page 363](#)
- 378. [Page 364](#)
- 379. [Page RE-1](#)
- 380. [Page RE-2](#)

381. [Page RE-3](#)
382. [Page RE-4](#)
383. [Page RE-5](#)
384. [Page RE-6](#)
385. [Page RE-7](#)
386. [Page RE-8](#)
387. [Page RE-9](#)
388. [Page RE-10](#)
389. [Page RE-11](#)
390. [Page RE-12](#)
391. [Page RE-13](#)
392. [Page RE-14](#)
393. [Page RE-15](#)

Guide

1. [Table of Contents and Preface](#)
2. [Cover Page](#)
3. [Title Page](#)
4. [Copyright Information](#)
5. [Dedication](#)
6. [About the Authors](#)
7. [Brief Contents](#)
8. [Contents](#)
9. [Preface](#)
10. [Features of This Canadian Edition](#)
11. [Organization & Other changes](#)
12. [Award-Winning Technology](#)
13. [Acknowledgments](#)
14. [Connect](#)
15. [Chapter 1: Scientific Understanding of Behaviour](#)
16. [Chapter 1 Introduction](#)
17. [Why Study Research Methods?](#)
18. [Methods of Acquiring Knowledge](#)
19. [Intuition](#)
20. [Authority](#)
21. [LO2 The Scientific Method: Be Skeptical, Seek Empirical Data](#)

22. [Science as a Way to Ask and Answer Questions](#)
23. [Goals of Scientific Research in Psychology](#)
24. [Describing Behaviour](#)
25. [Predicting Behaviour](#)
26. [Determining the Causes of Behaviour](#)
27. [Explaining Behaviour](#)
28. [Basic and Applied Research](#)
29. [Basic Research](#)
30. [Applied Research](#)
31. [Integrating Basic and Applied Research](#)
32. [Study Terms](#)
33. [Review Questions](#)
34. [Deepen Your Understanding](#)
35. [Chapter 2: Where to Start](#)
36. [Chapter 2 Introduction](#)
37. [Where Do Research Ideas Come From?](#)
38. [Questioning Common Assumptions](#)
39. [Observation of the World around Us](#)
40. [Practical Problems](#)
41. [LO2 Theories](#)
42. [Past Research](#)
43. [How Do We Find Out What is Already Known?](#)
44. [What to Expect in a Research Article](#)
45. [Other Types of Articles: Literature Reviews and Meta-analyses](#)
46. [Reading Articles](#)
47. [LO4 Where Are These Articles Published? An Orientation to Journals and Finding Articles](#)
48. [Developing Hypotheses and Predictions](#)
49. [Study Terms](#)
50. [Review Questions](#)
51. [Deepen Your Understanding](#)
52. [Chapter 3: Ethical Research](#)
53. [Chapter 3 Introduction](#)
54. [Were Milgram's Obedience Experiments Ethical?](#)
55. [Ethical Research in Canada](#)
56. [The Tri-Council and Its Policy Statement](#)
57. [Historical, Legal, and International Context](#)

58. [LO1 Core Principles Guiding Research with Human Participants](#)
59. [Designing Research to Uphold the Core Principles](#)
60. [LO2 Promote Concern for Welfare by Minimizing Risks and Maximizing Benefits](#)
61. [LO3 Promote Respect for Persons through Informed Consent](#)
62. [Promote Justice by Involving People Equitably in Research](#)
63. [Evaluating the Ethics of Research with Human Participants](#)
64. [Monitoring Ethical Standards at Each Institution](#)
65. [LO6 Exempt Research](#)
66. [Minimal Risk Research](#)
67. [Greater Than Minimal Risk Research](#)
68. [Ethics and Animal Research](#)
69. [Professional Ethics in Academic Life](#)
70. [Ethics Codes of the APA and CPA](#)
71. [LO8 Scientific Misconduct and Publication Ethics](#)
72. [Plagiarism and the Integrity of Academic Communication](#)
73. [Study Terms](#)
74. [Review Questions](#)
75. [Deepen Your Understanding](#)
76. [Chapter 4: Research Design Fundamentals](#)
77. [Chapter 4 Introduction](#)
78. [Introduction to Basic Research Design](#)
79. [Variables](#)
80. [LO1 Two Basic Research Designs](#)
81. [LO2 Operationally Defining Variables: Turning Hypotheses into Predictions](#)
82. [Non-experimental Method](#)
83. [LO4 Relationships between Variables](#)
84. [Interpreting the Results of Non-experimental designs](#)
85. [Experimental Method](#)
86. [LO6 Designing Experiments That Allow for Causal Inferences](#)
87. [Choosing a Method: Advantages of Multiple Methods](#)
88. [Artificiality of Experiments](#)
89. [Ethical and Practical Considerations](#)
90. [Describing Behaviour](#)
91. [Predicting Future Behaviour](#)
92. [Advantages of Multiple Methods](#)

- 93. [Study Terms](#)
- 94. [Review Questions](#)
- 95. [Deepen Your Understanding](#)
- 96. [Chapter 5: Measurement](#)
- 97. [Chapter 5 Introduction](#)
- 98. [Self-Report Measures](#)
- 99. [Reliability](#)
- 100. [Test-Retest Reliability](#)
- 101. [Internal Consistency Reliability](#)
- 102. [Inter-rater Reliability](#)
- 103. [LO2 Reliability and Accuracy of Measures](#)
- 104. [Validity of Measures](#)
- 105. [Indicators of Construct Validity](#)
- 106. [Reactivity of Measures](#)
- 107. [Variables and Measurement Scales](#)
- 108. [Nominal Scales](#)
- 109. [Ordinal Scales](#)
- 110. [Interval Scales](#)
- 111. [Ratio Scales](#)
- 112. [The Importance of the Measurement Scales](#)
- 113. [Study Terms](#)
- 114. [Review Questions](#)
- 115. [Deepen Your Understanding](#)
- 116. [Chapter 6: Observational Methods](#)
- 117. [Chapter 6 Introduction](#)
- 118. [Quantitative and Qualitative Approaches](#)
- 119. [Naturalistic Observation](#)
- 120. [Issues in Naturalistic Observation](#)
- 121. [Systematic Observation](#)
- 122. [Coding Schemes](#)
- 123. [Issues in Systematic Observation](#)
- 124. [Case Studies](#)
- 125. [Archival Research](#)
- 126. [Census Data or Statistical Records](#)
- 127. [Survey Archives](#)
- 128. [Written Records and Mass Media](#)
- 129. [Working with Archival Data: Content Analysis and Interpretation](#)

- 130. [Study Terms](#)
- 131. [Review Questions](#)
- 132. [Deepen Your Understanding](#)
- 133. [Chapter 7: Survey Research: Asking People about Themselves](#)
- 134. [Chapter 7 Introduction](#)
- 135. [Why Conduct Surveys?](#)
- 136. [Response Bias in Survey Research](#)
- 137. [Constructing Good Questions](#)
- 138. [Defining the Research Objectives](#)
- 139. [Question Wording](#)
- 140. [Responses to Questions: What Kind of Data Are You Seeking?](#)
- 141. [Closed- versus Open-Ended Questions](#)
- 142. [Rating Scales for Closed-Ended Questions](#)
- 143. [Finalizing the Questionnaire](#)
- 144. [Formatting the Questionnaire](#)
- 145. [Refining Questions](#)
- 146. [Administering Surveys](#)
- 147. [Questionnaires](#)
- 148. [Interviews](#)
- 149. [Interpreting Survey Results: Consider the Sample](#)
- 150. [Population and Samples](#)
- 151. [LO5 For More Precise Estimates, Use a Larger Sample](#)
- 152. [LO6 To Describe a Specific Population, Sample Thoroughly](#)
- 153. [Sampling Techniques](#)
- 154. [Probability Sampling](#)
- 155. [Non-probability Sampling](#)
- 156. [Reasons for Using Convenience Samples](#)
- 157. [Study Terms](#)
- 158. [Review Questions](#)
- 159. [Deepen Your Understanding](#)
- 160. [Chapter 8: Experimental Design](#)
- 161. [Chapter 8 Introduction](#)
- 162. [Confounding and Internal Validity](#)
- 163. [Planning a Basic Experiment](#)
- 164. [Between-Subjects Experiments](#)
- 165. [LO3 Pretest-Posttest Design](#)
- 166. [LO4 Matched Pairs Design](#)

167. [Within-Subjects Experiments](#)
168. [LO5 Advantages and Disadvantages of the Within-Subjects Design](#)
169. [Counterbalancing](#)
170. [Time Interval between Treatments](#)
171. [Choosing between Between-Subjects and Within-Subjects Designs](#)
172. [Study Terms](#)
173. [Review Questions](#)
174. [Deepen Your Understanding](#)
175. [Chapter 9: Conducting Studies](#)
176. [Chapter 9 Introduction](#)
177. [Finalizing a Study Design](#)
178. [LO1 Options for Manipulating the Independent Variable in Experiments](#)
179. [Additional Considerations when Manipulating the Independent Variable](#)
180. [LO2 Options for Measuring Variables](#)
181. [LO3 Additional Considerations when Measuring Variables](#)
182. [LO4 Setting the Stage](#)
183. [Advanced Considerations for Ensuring Control](#)
184. [Controlling for Participant Expectations](#)
185. [Controlling for Experimenter Expectations](#)
186. [Seeking Ethics Approval](#)
187. [Selecting Research Participants](#)
188. [Planning the Debriefing](#)
189. [Collecting Data](#)
190. [Pilot Studies](#)
191. [Researcher Commitments](#)
192. [What Comes Next?](#)
193. [Analyzing and Interpreting Results](#)
194. [Communicating Research to Others](#)
195. [Study Terms](#)
196. [Review Questions](#)
197. [Deepen Your Understanding](#)
198. [Chapter 10: Research Designs for Special Circumstances](#)
199. [Chapter 10 Introduction](#)
200. [Program Evaluation](#)
201. [Quasi-Experimental Designs](#)

- 202. [LO2 One-Group Posttest-Only Design](#)
- 203. [One-Group Pretest-Posttest Design](#)
- 204. [LO3 Threats to Internal Validity](#)
- 205. [LO4 Non-equivalent Control Group Design](#)
- 206. [Non-equivalent Control Group Pretest-Posttest Design](#)
- 207. [LO5 Interrupted Time Series Design](#)
- 208. [Control Series Design](#)
- 209. [Summing up Quasi-Experimental Designs](#)
- 210. [Single Case Experimental Designs](#)
- 211. [Reversal Designs](#)
- 212. [Multiple Baseline Designs](#)
- 213. [Replications in Single Case Designs](#)
- 214. [Developmental Research Designs](#)
- 215. [Longitudinal Method](#)
- 216. [Cross-Sectional Method](#)
- 217. [Comparing Longitudinal and Cross-Sectional Methods](#)
- 218. [Sequential Method](#)
- 219. [Study Terms](#)
- 220. [Review Questions](#)
- 221. [Deepen Your Understanding](#)
- 222. [Chapter 11: Complex Experimental Designs](#)
- 223. [Chapter 11 Introduction](#)
- 224. [An Independent Variable with More Than Two Levels](#)
- 225. [An Experiment with More Than One Independent Variable: Factorial Designs](#)
- 226. [LO2 Interpreting Factorial Designs](#)
- 227. [Interactions Illuminate Moderator Variables](#)
- 228. [LO3 Depicting Possible Outcomes of a \$2 \times 2\$ Factorial Design Using Tables and Graphs](#)
- 229. [LO4 Breaking Down Interactions into Simple Main Effects](#)
- 230. [Variations on \$2 \times 2\$ Factorial Designs](#)
- 231. [LO5 Factorial Designs with Manipulated and Non-manipulated Variables](#)
- 232. [LO6 Assignment Procedures and Sample Size](#)
- 233. [Increasing the Complexity of Factorial Designs](#)
- 234. [Beyond Two Levels per Independent Variable](#)
- 235. [Beyond Two Independent Variables](#)

- 236. [Study Terms](#)
- 237. [Review Questions](#)
- 238. [Deepen Your Understanding](#)
- 239. [Chapter 12: Descriptive Statistics: Describing Variables and the Relations among Them](#)
- 240. [Chapter 12 Introduction](#)
- 241. [Revisiting Scales of Measurement](#)
- 242. [Describing Each Variable](#)
- 243. [LO1 Graphing Frequency Distributions](#)
- 244. [LO2 Descriptive Statistics](#)
- 245. [Describing Relationships Involving Nominal Variables](#)
- 246. [Comparing Groups of Participants](#)
- 247. [Graphing Nominal Data](#)
- 248. [LO4 Describing Effect-Size between Two Groups](#)
- 249. [Describing Relationships among Continuous Variables: Correlating Two Variables](#)
- 250. [Interpreting the Pearson r Correlation Coefficient](#)
- 251. [Scatterplots](#)
- 252. [Important Considerations](#)
- 253. [Correlation Coefficients as Effect-Sizes](#)
- 254. [Describing Relationships among Continuous Variables: Increasing Complexity](#)
- 255. [LO6 The Regression Equation](#)
- 256. [Multiple Correlation and Multiple Regression](#)
- 257. [Integrating Results from Different Analyses](#)
- 258. [LO7 Partial Correlation and the Third-Variable Problem](#)
- 259. [Advanced Modelling Techniques](#)
- 260. [Combining Descriptive and Inferential Statistics](#)
- 261. [Study Terms](#)
- 262. [Review Questions](#)
- 263. [Deepen Your Understanding](#)
- 264. [Chapter 13: Inferential Statistics: Making Inferences about Populations Based on Our Samples](#)
- 265. [Chapter 13 Introduction](#)
- 266. [Inferential Statistics: Using Samples to Make Inferences about Populations](#)
- 267. [Inferential Statistics: Ruling Out Chance](#)

- 268. [Statistical Significance: An Overview](#)
- 269. [Null and Research Hypotheses](#)
- 270. [Probability and Sampling Distributions](#)
- 271. [Probability: The Case of Mind Reading](#)
- 272. [LO3 Sampling Distributions](#)
- 273. [Sample Size](#)
- 274. [How “Unlikely” Is Enough? Choosing a STATISTICAL Significance Level \(Alpha\)](#)
- 275. [Example Statistical Tests](#)
- 276. [LO4 Thet-Test: Comparing Two Means](#)
- 277. [LO5 TheFTest: Used When Comparing Three or More Group Means](#)
- 278. [Statistical Significance of a Pearson r Correlation Coefficient](#)
- 279. [We Made a Decision about the Null Hypothesis, but We Might Be Wrong! Investigating Type I and Type II Errors](#)
- 280. [Correct Decisions](#)
- 281. [Type I Errors](#)
- 282. [Type II Errors](#)
- 283. [The Everyday Context of Type I and Type II Errors](#)
- 284. [Type I and Type II Errors in the Published Research Literature](#)
- 285. [Interpreting Statistically Non-significant Results](#)
- 286. [Choosing a Sample Size: Power Analysis](#)
- 287. [Analyzing Data Using Statistics Software](#)
- 288. [Selecting the Appropriate Statistical Test](#)
- 289. [Research Studying Two Variables](#)
- 290. [Research with Multiple Independent or Predictor Variables](#)
- 291. [Integrating Descriptive and Inferential Statistics](#)
- 292. [Effect-Size](#)
- 293. [Confidence Intervals and Statistical Significance](#)
- 294. [Conclusion Validity](#)
- 295. [The Importance of Replications](#)
- 296. [Study Terms](#)
- 297. [Review Questions](#)
- 298. [Deepen Your Understanding](#)
- 299. [Chapter 14: Generalizing Results](#)
- 300. [Chapter 14 Introduction](#)
- 301. [Challenges to Generalizing Results](#)
- 302. [LO1 Can Results Generalize to Other Populations?](#)

- 303. [LO2 Can Results Generalize beyond the Specific Study Situation?](#)
- 304. [Solutions to Generalizing Results](#)
- 305. [LO3 Replicate the Study](#)
- 306. [Consider Different Populations](#)
- 307. [LO5 Rely on Multiple Studies to Draw Conclusions: Literature Reviews and Meta-analyses](#)
- 308. [Generalizing Your Knowledge beyond This Book](#)
- 309. [Recognize and Use Your New Knowledge](#)
- 310. [Stay Connected to Building a Better Psychological Science](#)
- 311. [Use Research to Improve Lives](#)
- 312. [Study Terms](#)
- 313. [Review Questions](#)
- 314. [Deepen Your Understanding](#)
- 315. [Appendix A: Writing Research Reports in APA Style](#)
- 316. [Appendix A Introduction](#)
- 317. [Writing Style and Word Choice](#)
- 318. [Clarity and Intended Audience](#)
- 319. [Paraphrase and Cite Past Research](#)
- 320. [Active versus Passive Voice](#)
- 321. [Avoiding Biased Language](#)
- 322. [Writing Each Section of the APA Style Research Report](#)
- 323. [Title Page](#)
- 324. [Abstract](#)
- 325. [Introduction](#)
- 326. [Method](#)
- 327. [Results](#)
- 328. [Discussion](#)
- 329. [References](#)
- 330. [Appendix](#)
- 331. [Author Note](#)
- 332. [Footnotes and Endnotes](#)
- 333. [Tables](#)
- 334. [Figures](#)
- 335. [Summary: Order of Pages](#)
- 336. [Formatting a Manuscript](#)
- 337. [Using Headings](#)
- 338. [Abbreviations](#)

- 339. [Reporting Numbers and Statistics](#)
- 340. [APA Style and Student Paper Formats](#)
- 341. [Citing and Referencing Sources](#)
- 342. [Citing Sources in the Body of the Report](#)
- 343. [Reference List Style](#)
- 344. [Reference Formats for Electronic Sources](#)
- 345. [Conference Presentations](#)
- 346. [Paper Presentations](#)
- 347. [Poster Sessions](#)
- 348. [Appendix A Sample Paper](#)
- 349. [Appendix B: Statistical Tests](#)
- 350. [Appendix B Introduction](#)
- 351. [Descriptive Statistics](#)
- 352. [Measures of Central Tendency](#)
- 353. [Measures of Variability](#)
- 354. [Pearson Product-Moment Correlation Coefficient](#)
- 355. [Additional Statistical Significance Tests](#)
- 356. [Significance of Correlation Coefficient r](#)
- 357. [Chi-Square \(\$\chi^2\$ \)](#)
- 358. [Analysis of Variance \(F Test\): Overview](#)
- 359. [Analysis of Variance: One Independent Variable, Between-Subjects Design](#)
- 360. [Analysis of Variance: Two Independent Variables, Between-Subjects Design](#)
- 361. [Analysis of Variance: One Independent Variable, Within-subjects Design](#)
- 362. [Analysis of Variance: Conclusion](#)
- 363. [Effect Size](#)
- 364. [Effect Size as Strength of Association](#)
- 365. [Effect Size as Proportion of Variance Explained](#)
- 366. [Effect Size as a Standardized Difference between Means](#)
- 367. [Appendix C: Statistical Tables](#)
- 368. [Appendix C Introduction](#)
- 369. [Appendix D: How to Conduct a PsycINFO Search](#)
- 370. [Appendix D Introduction](#)
- 371. [The User Screen](#)
- 372. [Specifying Search Terms and Finding Results](#)

- 373. [Combining Search Terms and Narrowing Results](#)
- 374. [Saving Results](#)
- 375. [Appendix E: Constructing a Latin Square](#)
- 376. [Appendix E Introduction](#)
- 377. [References](#)
- 378. [References](#)
- 379. [Sources](#)
- 380. [Accessibility Content: Text Alternatives for Images](#)
- 381. [Figure 1.1 Text Alternative \(Chapter 1\)](#)
- 382. [Figure 2.1 Text Alternative \(Chapter 2\)](#)
- 383. [Figure 3.1 Text Alternative \(Chapter 3\)](#)
- 384. [Figure 3.3 Text Alternative \(Chapter 3\)](#)
- 385. [Figure 4.1 Text Alternative \(Chapter 4\)](#)
- 386. [Figure 4.2 Text Alternative \(Chapter 4\)](#)
- 387. [Figure 4.3 Text Alternative \(Chapter 4\)](#)
- 388. [Figure 4.4 Text Alternative \(Chapter 4\)](#)
- 389. [Figure 4.5 Text Alternative \(Chapter 4\)](#)
- 390. [Figure 5.1 Text Alternative \(Chapter 5\)](#)
- 391. [Figure 5.2 Text Alternative \(Chapter 5\)](#)
- 392. [Figure 5.3 Text Alternative \(Chapter 5\)](#)
- 393. [A Flow Chart Planning a Between-Subjects Experiment Text Alternative \(Chapter 8\)](#)
- 394. [A Flow Chart Presented Listen then Read Text Alternative \(Chapter 8\)](#)
- 395. [Figure 10.1 Text Alternative \(Chapter 10\)](#)
- 396. [A Flow Chart Concerning Independent and Dependent Variables Text Alternative \(Chapter 10\)](#)
- 397. [A Flow Chart One-Group Pretest-Posttest Design Text Alternative \(Chapter 10\)](#)
- 398. [Two Flow Charts Non-Equivalent Control Group Design Text Alternative \(Chapter 10\)](#)
- 399. [Two Flow Charts Non-Equivalent Control Group Pretest-Posttest Design Text Alternative \(Chapter 10\)](#)
- 400. [Figure 10.2 Text Alternative \(Chapter 10\)](#)
- 401. [Figure 10.3 Text Alternative \(Chapter 10\)](#)
- 402. [Figure 10.4 Text Alternative \(Chapter 10\)](#)
- 403. [Figure 10.5 Text Alternative \(Chapter 10\)](#)
- 404. [Figure 11.1 Text Alternative \(Chapter 11\)](#)

- 405. [Figure 11.2 Text Alternative \(Chapter 11\)](#).
- 406. [Figure 11.3 Text Alternative \(Chapter 11\)](#).
- 407. [Figure 11.4 Text Alternative \(Chapter 11\)](#).
- 408. [Figure 11.5 Text Alternative \(Chapter 11\)](#).
- 409. [Figure 11.6 Text Alternative \(Chapter 11\)](#).
- 410. [Figure 11.7 Text Alternative \(Chapter 11\)](#).
- 411. [Figure 12.1 Text Alternative \(Chapter 12\)](#).
- 412. [Figure 12.2 Text Alternative \(Chapter 12\)](#).
- 413. [Figure 12.3 Text Alternative \(Chapter 12\)](#).
- 414. [Figure 12.4 Text Alternative \(Chapter 12\)](#).
- 415. [Figure 12.5 Text Alternative \(Chapter 12\)](#).
- 416. [Figure 12.6 Text Alternative \(Chapter 12\)](#).
- 417. [Figure 12.7 Text Alternative \(Chapter 12\)](#).
- 418. [Figure 12.8 Text Alternative \(Chapter 12\)](#).
- 419. [Figure 12.9 Text Alternative \(Chapter 12\)](#).
- 420. [Figure 13.1 Text Alternative \(Chapter 13\)](#).
- 421. [Figure 13.2 Text Alternative \(Chapter 13\)](#).
- 422. [Figure 13.3 Text Alternative \(Chapter 13\)](#).
- 423. [Figure 13.4 Text Alternative \(Chapter 13\)](#).
- 424. [Figure 14.1 Text Alternative \(Chapter 14\)](#).
- 425. [Figure 14.2 Text Alternative \(Chapter 14\)](#).
- 426. [Figure A.1 Text Alternative \(Appendix A\)](#).
- 427. [Figure 1 Text Alternative \(Appendix A\)](#).
- 428. [Figure D.1 Text Alternative \(Appendix D\)](#).

Glossary

abstract

The section of a research report at the very beginning that briefly summarizes the entire study or studies.

alpha level

In NHST, the threshold probability below which a test statistic is deemed to be unlikely to have originated from the sampling distribution (set at .05, by convention). Probability values (*p*-values) below this threshold are viewed as statistically significant.

alternative explanation

Part of causal inference; a potential alternative cause of an observed relationship between variables.

analysis of variance (ANOVA)

See F test.

anonymous

Protecting the identity of participants by making them unidentifiable based on the data collected; with anonymous data, it is impossible to identify which participant provided what data.

applied research

Research conducted to address practical problems and propose potential solutions.

archival research

The use of existing sources of information for research, such as census data, archived survey data, and other forms of preserved written records.

authority

Any source of power or control (e.g., news media, books, government officials, or religious figures). We often defer to authorities and accept their ideas and recommendations unthinkingly.

bar graph

A graph using bars to depict frequencies of responses, percentages, or means in two or more groups.

baseline

A form of control condition in which participant behaviour is measured during a control period, before introduction of the manipulation.

basic research

Research that attempts to answer fundamental questions about the nature of behaviour.

behavioural measure

An operationalization of a variable that involves directly observing and precisely recording behaviour.

case study

An in-depth analysis of a single person or setting that often includes detailed descriptive accounts of behaviour, past history, and other relevant factors.

ceiling effect

The failure of a measure to detect difference because it was too easy and everyone does well (*see also floor effect*).

cell

A single entry in a table, sometimes used to refer to one condition in an experiment, or to a combination of conditions in a factorial design.

central tendency

A single number or value that attempts to summarize all of the data, describing the typical score or where most of the scores fall.

citations

Names and dates referencing another publication that appear in the body of a text. These serve to properly attribute ideas and results to the authors being cited rather than the current paper's author. Citations refer readers to the corresponding entry in the *references* section for full details regarding the publication.

closed-ended questions

Questions that offer respondents a limited number of response options.

An example of this is a multiple choice question.

cluster sampling

A method of sampling in which clusters of individuals are identified. Clusters are sampled, and then all individuals in each cluster are included in the sample.

coding scheme

A set of rules used to categorize observations during systematic observation.

Cohen's d

An effect-size estimate that is the standardized mean difference in scores between two groups.

cohort effects

A cohort is a group of people born at about the same time and exposed to the same societal events; cohort effects are confounded with age in a cross-sectional study.

concealed observation

A type of naturalistic observation in which those being studied are not aware of the researcher's presence.

conceptual replication

An attempted replication of past research using different procedures than the original study, such as a different dependent measure or different manipulation.

concern for welfare

The ethical principle that research should maximize benefits and minimize harm.

conclusion validity

The extent to which the conclusions about the relationships among variables, based on our analysis of the data, are correct.

concurrent validity

A type of construct validity that examines whether the measure can predict a criterion measured at the same time, rather than some criterion behaviour in the future (as is the case for predictive validity).

confederate

A person posing as a participant in an experiment who is actually a collaborator of the experimenter.

confidence interval

In practice, the confidence interval serves as a range of plausible values that are likely to be observed, if the study were to be repeated. This is a quantification of our uncertainty around our estimates, with wider confidence intervals indicating greater uncertainty.

confidential

The ethical principle that information is kept private, with disclosure limited to the minimum number of people necessary.

confounding variables

Variables that are not of interest, but are uncontrolled and impossible to separate from the variable of interest, and that can also explain the pattern of results. These confounds make interpreting the results difficult or impossible, as we do not know what is responsible for an effect or association.

construct validity

The degree to which a measure accurately measures the theoretical construct it is designed to measure.

content analysis

Systematic analysis of the content of written records.

content validity

A form of construct validity evaluated by comparing the content of the measure to the theoretical definition of the construct, ensuring that all aspects of the construct are measured and no extraneous elements are also measured.

contrast effect

In a within-subjects design, occurs when participants' responses in a later condition are affected by a particular experience they had in an earlier condition.

control series design

An extension of the interrupted time series quasi-experimental design in which there is a comparison or control group.

convenience sampling

A type of non-probability sampling that involves selecting participants in a haphazard manner, usually on the basis of availability, and without regard for the representativeness of the sample.

convergent validity

An aspect of construct validity assessed by examining the extent to which scores on the measure are related to other measures of the same or similar constructs.

correlation coefficient

A statistic that describes how strongly two variables are related to one another, the degree to which they covary.

counterbalancing

A method of controlling for order effects in a within-subjects design, by either including all possible orders of presentation for conditions or

randomly determining the order for each participant.

covariation of cause and effect

One of the criteria for making an appropriate causal inference; observing that a change in one variable is accompanied by a change in the other variable.

criterion variable

The outcome variable that is being predicted in a regression analysis.

Cronbach's alpha

An indicator of internal consistency reliability assessed by examining the average correlation of each item in a measure with every other question.

cross-sectional method

A developmental research method in which people of different ages are studied at a single point in time; conceptually similar to a between-subjects design.

curvilinear relationship

A relationship between two variables in which increases in the values of the first variable are accompanied by both increases and decreases in the values of the second variable, resulting in a curved line depicting this relationship rather than a straight line.

debriefing

An explanation of the purpose of the research, given to participants following their participation.

deception

Any time a researcher misleads participants into believing something about a study that is not true.

degrees of freedom (df)

A concept used in tests of statistical significance; the number of observations that are free to vary to produce a known outcome.

demand characteristics

Cues embedded in a study that inform the participant how he or she is expected to behave.

dependent variable

The variable measured in an experiment that represents some outcome of interest, which is hypothesized to be affected by the causal variable (i.e., the independent variable).

descriptive statistics

Statistics that describe and summarize the data collected; these include measures of central tendency (e.g., mean), variability (e.g., standard deviation), and covariation (e.g., Pearson correlation).

direct replication

An attempted replication of a study following the same procedures that were used in the original research as closely as possible.

discriminant validity

An aspect of construct validity in which scores on a measure are not related to scores on conceptually unrelated measures; also known as *divergent validity*.

discussion

The section of a research report in which the researcher considers the research results from various perspectives.

double-blind procedure

An experiment in which neither the experimenter nor the participant knows to which condition the participant has been assigned.

effect-size

The magnitude of an effect observed, either the extent to which two variables are associated or the size of the difference in scores between groups.

empiricism

Gaining knowledge through systematic observations of the world.

error variance

The variability of each score relative to its group mean. This is random variability that is not attributed to changes in the independent variable.

ethics codes

Guidelines for ethical conduct developed by professional societies.

exempt research

Research in which there is absolutely no risk to participants, and is thus exempt from REB review.

experimental control

This quality of good experimental design is achieved when only the independent variable of interest varies across conditions, with all other features kept the same across conditions.

experimental method

A method that tries to determine if variables are causally related, by manipulating one variable (the independent variable), controlling all other variables, and then measuring the effect on some outcome (the dependent variable).

experimenter bias

Any intentional or unintentional influence that the experimenter exerts on participants to confirm the hypothesis; also called *experimenter expectancy effects*.

external validity

The degree to which study results, based on a sample, may be generalized to the population from which the sample was drawn or other populations.

F test

An NHST test for whether two or more means differ in the population; also known as an *analysis of variance (ANOVA)*. The *F* statistic is the ratio of systematic variance to error variance.

face validity

The degree to which a measure appears to measure the intended variable.

factorial design

An experiment with more than one independent variable (also called a *factor*), with each factor having at least two levels (i.e., two conditions).

falsifiability

The principle that a good scientific idea or theory should be capable of being shown to be false when tested using scientific methods.

falsifiable

Capable of being shown to be false; a good scientific idea or theory should be falsifiable.

fatigue effect

When participants perform worse over the course of a study simply because of diminished effort or fewer resources due to the passage of time; particularly problematic in within-subjects designs.

field experiment

An experiment that is conducted in a natural setting, out in the real world, rather than in a laboratory.

filler items

Items included in a questionnaire measure to help disguise the true purpose of the measure.

floor effect

The failure of a measure to detect a difference because it was too difficult (*see also ceiling effect*).

focus group

A qualitative method of data collection in which six to ten people are interviewed together about a particular topic.

fraud

The intentional misrepresentation of any aspect of research, including the presentation of results that are misleading or based on faulty data.

frequency distribution

A representation of how often each score was observed, arranged from lowest to highest score.

frequency polygons

Graphs of frequencies for continuous variables, in which the frequency of each score is plotted on the vertical axis and these points are connected by straight lines.

goals of scientific research

The four main goals of scientific research are: (1) to describe behaviour, (2) to predict behaviour, (3) to determine the causes of behaviour, and (4) to understand or explain behaviour.

graphic rating scale

A type of closed-ended response scale option where two words appear on either side of a solid line. Participants place a mark on the line indicating their relative preference for one or the other word, which is then measured in terms of distance from one end.

histogram

A type of bar graph used when the variable on the x -axis is continuous, with each bar touching the adjacent bars (unlike in typical bar graphs).

history effects

Threats to internal validity, in which outside events not part of the manipulation influence the dependent variable, providing an alternative explanation for the results.

between-subjects design

An experimental design in which different participants are assigned to each level or condition of the independent variable. Also called an *independent groups design*.

independent variable

The variable in an experiment that is manipulated by the researcher, in order to observe its effect on some outcome variable (i.e., the dependent variable).

inferential statistics

Statistics that estimate whether the results observed based on sample data are generalizable to the population from which that sample was drawn.

informed consent

The ethical principle that potential participants be informed in advance of all aspects of the research that might influence their decision to participate.

instrument decay

A threat to internal validity in which the characteristics of the measurement instrument changes over time, providing an alternative explanation for the results observed.

interaction

When the effect of one independent variable on the dependent variable depends on the level of another independent variable.

internal consistency reliability

A form of reliability assessing the degree to which the items in a scale are consistent in measuring the same construct or variable.

internal validity

A quality experiments, referring to the certainty with which it can support causal inferences; the degree to which the outcomes observed can be attributed to the manipulated independent variable, rather than some alternative explanation.

inter-rater reliability

An indicator of reliability that examines the degree to which two or more raters agree, having the same or similar judgments for a set of stimuli.

interrupted time series design

A quasi-experimental design in which a treatment is investigated by examining a series of measurements made over an extended time

period, both before and after the treatment is introduced.

interval scale

A scale of measurement in which the intervals between numbers on the scale are all equal in size.

interviewer bias

Intentional or unintentional influence on a respondent exerted by an interviewer, which might encourage certain responses consistent with the interviewer's expectations.

introduction

The section of a research report in which the researcher outlines the problem that has been investigated.

intuition

Relying upon anecdote, experience, or judgment to make sense of the world, without adopting a critical or questioning mindset.

IV × PV design

A factorial design that includes both an experimental independent variable (IV) and a non-experimental participant variable (PV).

justice

The ethical principle that all individuals and groups should have fair and equal access to the benefits of research participation, and bear its potential risks equally.

Latin square design

A technique to control for order effects without having all possible orders.

levels

The operationalization of the independent variable in an experiment, often referred to as the conditions or groups.

literature review

A narrative summary of the past research conducted on a particular topic.

longitudinal method

A developmental research method in which the same people are observed repeatedly as they grow older; conceptually similar to a within-subjects design.

main effect

The direct effect of an independent variable on a dependent variable, ignoring any interaction with other variables.

manipulation check

A measure used to determine whether the manipulation of the independent variable has had its intended effect on a participant.

manipulation strength

The degree to which levels of an independent variable differ from each other. In a weak manipulation, conditions are subtly different. In a strong manipulation, conditions are maximally different.

marginal mean

In a factorial design, the average score of all participants in one condition of one independent variable, collapsing across all other variables.

matched pairs design

In a between-subjects experiment, a method of assigning participants to conditions in which pairs of participants are first matched on some characteristic and then each member of the pair is randomly assigned to a condition.

maturation effects

Threats to internal validity, in which any naturally occurring change within individuals that occurs over time could provide an alternative explanation for the results.

mean

A measure of central tendency, obtained by summing scores and then dividing this sum by the number of scores.

measurement error

Anything that contributes to the score on a measure that is not based on the true score. In other words, this is responsible for the degree to which a score on a measure deviates from the true value of the measured variable.

median

A measure of central tendency defined as the middle score in a distribution that divides the distribution in half (or an average of the two middle scores).

mediating variable

A psychological process that occurs between an event and a behavioural response.

meta-analysis

A statistical procedure for combining the results of many past studies in order to provide an estimate of the effect-size for this general phenomenon.

method

The section of a research report providing information about exactly how the study was conducted, including any details necessary for the reader to replicate the study.

minimal risk research

Research that involves no greater risks to participants than they would typically encounter in their daily lives.

mixed factorial design

A factorial experimental design that includes both between-subjects and within-subjects variables.

mode

A measure of central tendency; the most frequent score in a distribution of scores.

moderator variable

A third variable that influences the relationship between an independent variable and a dependent variable. In a factorial design, the effect of moderator variables are revealed as interactions.

selective attrition

The loss of participants by way of individuals choosing to drop out of an experiment. Selective attrition is a threat to internal validity when participant dropout results in a difference between conditions on some participant characteristic, causing a confound.

multiple baseline design

Observing behaviour before and after a manipulation under multiple circumstances (across different individuals, different behaviours, or different settings).

multiple correlation

A correlation between a combined set of predictor variables and one criterion variable.

multiple regression

An extension of the correlation technique that models the extent to which one or more predictor variables are related to one criterion variable.

naturalistic observation

Systematic observations made in a natural setting in the real world; sometimes called *field observation* and useful for generating rich descriptions of phenomena.

negative linear relationship

A relationship between two variables in which increases in the values of the first variable are accompanied by decreases in the values of the second variable.

nominal scale

A scale of measurement with two or more categories that have no numerical properties (e.g., no "less than" or "greater than"); also known as *categorical variables*.

non-equivalent control group design

A quasi-experimental design in which the groups of participants in the different conditions are not equivalent (e.g., studying naturally occurring groups), and there is no pretest.

non-equivalent control group pretest-posttest design

A quasi-experimental design in which non-equivalent groups are used, but a pretest allows assessment of equivalency and pretest-posttest changes.

non-experimental method

Measuring variables to determine whether they are related to one another; also called the *correlational method*.

non-probability sampling

A type of sampling procedure in which one cannot specify, or does not know, the probability that any member of the population will be included in the sample.

null hypothesis

The hypothesis that there is no effect in the population (e.g., the experimental groups have equal means, the variables are not related), that any effect observed in our sample is due to random error and does not represent the population; typically contrasted with the research hypothesis.

one-group posttest-only design

A quasi-experimental design that has no control group and no pretest comparison; a very poor design in terms of internal validity.

one-group pretest-posttest design

A quasi-experimental design in which the effect of an independent variable is inferred from the pretest-posttest difference in a single group.

open-ended questions

Questions that allow respondents to answer in any way they wish, with no restrictions (e.g., an essay question).

operationalization

Definition of a variable that specifies the operation used to measure or manipulate it within a specific study; also known as an *operational definition*.

order effect

In a within-subjects design, the effect that the order of conditions has on the dependent variable.

ordinal scale

A scale of measurement in which the measurement categories form a rank order along a continuum.

outliers

Scores that are very different from the rest of the scores in a dataset (i.e., much smaller or much larger); also known as *extreme scores*.

panel study

In survey research, administering questions to the same people at two or more points in time. Also known as a *longitudinal design*.

parsimony

The scientific principle stating that if two theories are equally effective at explaining a phenomenon, the simpler of the two theories is preferable.

partial correlation

The correlation between two variables with the influence of a third variable statistically controlled.

participant observation

A type of naturalistic observation in which the researcher assumes an active role in the setting being researched. The researcher's purpose may or may not be concealed.

participant variable

A pre-existing characteristic or aspect of a person that is of interest.

Pearson correlation coefficient

A statistic indicating the strength of relationship between two variables, and whether this relationship is positive or negative. Appropriate for interval and ratio scale data, when the sample is drawn from a normally distributed population.

peer review

The process of judging the scientific merit of research through review by peers of the researcher—other scientists with the expertise to evaluate the research.

physiological measure

An operationalization of a variable that involves observing and recording a response from the body.

pie chart

A circular graph in which frequencies or percentages are represented as different “slices” of a pie.

pilot study

A small-scale study with a small sample, conducted prior to conducting an actual study, designed to test and refine procedures.

placebo group

In experiments, a control group given the expectation of improvement through treatment, in order to control for the psychological effects of receiving a treatment.

plagiarism

The intentional or unintentional use of another person’s work (words or ideas) without adequately indicating the source.

population

The group of people of interest to the researchers, from which a sample is typically drawn.

positive linear relationship

A relationship between two variables, in which increases in the values of the first variable are accompanied by increases in the values of the second variable.

posttest-only design

A true experimental design in which the dependent variable (posttest) is measured only once, after manipulation of the independent variable.

power

Within the NHST framework, the probability of correctly rejecting the null hypothesis using a particular statistical test.

practice effect

When participants perform better over the course of a study simply because they are more experienced with the tasks; particularly problematic in within-subjects designs.

prediction

A statement that makes a specific assertion about what is believed will occur.

predictive validity

An aspect of construct validity that involves examining if a measure can predict a theoretically relevant future behaviour or criterion; also known as *criterion validity*.

predictor variable

The variable used to predict changes in the criterion (or outcome) variable in a regression analysis.

pretest-posttest design

An experimental design in which the dependent variable is measured both before (pretest) and after (posttest) manipulation of the independent variable.

probability sampling

Type of sampling procedure in which one is able to specify the probability that any member of the population will be included in the sample.

probability

The likelihood that a given event (among a specific set of possible events) will occur.

program evaluation

Research designed to evaluate programs (e.g., social reforms, innovations) designed to produce changes or certain outcomes in a target population.

pseudoscience

Claims that are made with evidence designed to appear scientific, but this evidence is not based on the principles of the scientific method.

PsycINFO

The American Psychological Association's searchable database of journal publications from the 1800s to the present.

publication bias

The bias in the literature that emerges because statistically significant results are more likely to be published in scientific journals than statistically non-significant results.

purposive sampling

A type of convenience sampling procedure conducted to obtain predetermined types of individuals for the sample.

qualitative approach

An approach to research that emphasizes people's lived experiences in their own words, and the researcher's interpretation of those experiences.

quantitative approach

An approach to research that emphasizes scientific empiricism in design, data collection, and statistical analyses.

quasi-experimental designs

A study design that has many features of an experiment, but due to necessity lacks some aspects of a true experimental design (and so cannot support causal inferences).

quota sampling

A sampling procedure in which the sample is chosen to reflect the numerical composition of various subgroups in the population. A convenience sampling technique is used to obtain the sample.

random assignment

In an experiment, using chance to determine which participants end up in which conditions, in order to control for the effects of extraneous variables not of interest to the researcher.

random sample

When everyone in a given population is equally likely to have been selected to participate in the study, that sample is said to be random.

rating scale

Closed-ended response option that asks participants to indicate the degree to which they agree with a particular statement.

ratio scale

A numeric scale of measurement, with equal intervals, in which there is a meaningful zero point, indicating total absence of the variable being measured (e.g., weight, duration, reaction time). As a result, ratios of numbers on the scale can be formed.

reactivity

When the act of measuring or observing something changes it. For example, if people know they are being observed, they may change their behaviour and not act as they normally would.

references

A list of all sources cited in the paper, providing all information required to locate the source.

regression equation

An equation that represents a line drawn to best fit a set of data points, allowing one to predict values of one variable based on another variable.

regression toward the mean

A statistical phenomenon in which extreme scores on a variable tend to be closer to the mean when a measurement is repeated; this change can be mistakenly attributed to some manipulation or intervention.

Regression to the mean is an alternative explanation for an observed change.

reliability

The degree to which a measure is consistent, providing a stable form of measurement.

within-subjects design

An experiment in which the same participants experience all levels of the independent variable (i.e., all conditions). Also called a *repeated measures design*.

replicate

To repeat a research study to determine whether the results can be duplicated.

replication

Repeating a study to see if one observes the same result, to increase confidence in that result or demonstrate that it is not systematically observable.

Research Ethics Board (REB)

An ethics review committee established to review research proposals within a university. The REB is composed of scientists, non-scientists, and legal experts.

research hypothesis

Within inferential statistics, the statement that some phenomenon exists within a population (e.g., a difference in means between

experimental groups, a relationship between variables); typically contrasted with the null hypothesis.

respect for persons

The ethical principle stating that all individuals should have the free and informed choice whether to participate in research.

response rate

The percentage of people selected for a sample who actually complete a survey.

response set

A pattern of responding to questions that is not related to the content of the questions themselves and thus provides inaccurate information.

response variable

A participant's reaction to some event.

restriction of range

When only a subset of a variable's possible values are sampled or observed, which can lead to misleading null or attenuated correlations.

results

The section of a research report in which the researcher presents the findings.

reversal design

An attempt to increase certainty in the effect of an intervention by demonstrating that the observed effects fade once an intervention is withdrawn. In a reversal design, also called a *withdrawal design*, the treatment is introduced after a baseline period and then withdrawn during a second baseline period. It may be extended by adding a second introduction of the treatment.

risk-benefit analysis

An evaluation of the potential hazards of conducting a study, weighed against the potential benefits to participants and to society.

sampling distribution

A frequency distribution of values obtained if a study was repeated an infinite number of times, using the exact same parameters. Used in inferential statistics to evaluate the likelihood of a given result, based on chance alone.

sampling error

The degree to which the estimate based on a sample deviates from the true population value.

sampling frame

The individuals or clusters of individuals in a population who might actually be selected for inclusion in the sample.

sampling

The process of choosing members of a population of interest to be included in a sample that is studied.

scatterplot

A graph of the relationship between two variables, in which pairs of scores are plotted on the x- and y-axes. This graph illustrates the relationship between two variables.

scientific skepticism

Not accepting something as true unthinkingly, but rather seeking out and evaluating the relevant evidence to shape our beliefs about what might be true.

selection differences

Differences in the type of participants who make up each group in a between-subjects experimental design.

self-report measure

An operationalization of a variable that involves asking people to explicitly indicate something about themselves (e.g., personality, behaviour, attitudes).

semantic differential scale

A type of closed-ended response where two words appear on either side of a series of dashed lines. Participants place a mark on the dash indicating their relative preference for one or the other word.

sensitivity

The ability of a measure to detect differences or changes.

sequential method

A combination of cross-sectional and longitudinal designs to study developmental research questions.

simple main effect

In a factorial design, the effect of one independent variable on the dependent variable, at one particular level of another independent variable.

simple random sampling

A sampling procedure in which each member of the population has an equal probability of being included in the sample.

single case experimental design

A research design in which the effect of the independent variable is assessed using data from a single participant.

single-blind procedure

An experiment in which participants do not know to which condition they have been assigned, but the experimenter does.

situational variable

A characteristic of some event or environment external to a participant, to which they are exposed.

squared correlation coefficient

A correlation coefficient that has been multiplied by itself, resulting in a value that reflects the proportion of variance shared between the two variables (i.e., the amount of variance explained in one variable by the other variable, and vice versa).

squared multiple correlation coefficient

The proportion of variance in the criterion that can be explained by the combined set of predictors for multiple correlation.

staged manipulations

Operationalizations of an independent variable that involve creating a complex situation. Participants then experience the situation and their responses are recorded. Deception is often used to conceal the fact that the situation is a ruse.

standard deviation

The average deviation of scores from the mean (the square root of the variance).

statistically significant

Within the NHST framework, observing that an outcome has a low probability of occurrence (typically defined as a p -value less than .05), assuming that the null hypothesis is correct.

straightforward manipulations

Operationalizations that involve manipulating the independent variable using instructions or other stimulus materials in a simple and obvious way.

stratified random sampling

A sampling procedure in which the population is divided into strata followed by random sampling from each stratum.

survey research

Questionnaires and interviews carefully designed to gather information from people about themselves.

systematic observation

Observation of one or more specific variables, usually made in a precisely defined setting.

systematic variance

Variability in a set of scores that is the result of the independent variable; statistically, the variability of each group mean from the grand mean of all participants.

t-test

An NHST statistic used to compare two means.

temporal precedence

One of the criteria for making an appropriate causal inference; the cause comes before the effect in time.

test-retest reliability

A form of reliability assessed by administering the same measure on two different occasions, and then calculating the correlation between the two different scores obtained.

testing effects

Threats to internal validity in which simply taking a pretest changes behaviour, without any effect of the independent variable.

theory

A framework that organizes and explains various findings related to a particular phenomenon, and in doing so generates new, testable hypotheses about that phenomenon.

third-variable problem

The possibility that a third, unmeasured, variable is responsible for the observed association between two other variables. Ruling out possible third variables, or controlling for them in experiments, helps us to uncover causal relationships.

Three Rs

In animal research, the principles of replacement (avoid using animals if possible), reduction (minimize the number of animals used), and refinement (modify procedures to minimize distress).

Tri-Council Policy Statement (TCPS)

In Canada, the official statement of ethical conduct for research involving humans; researchers and institutions are expected to adhere

to this document to receive federal research funds.

true score

An individual's actual level of the variable being measured, not to be confused with the score they get on the measure of that variable.

Type I error

An incorrect decision to reject the null hypothesis, when it is in fact true.

Type II error

An incorrect decision to accept the null hypothesis, when it is in fact false.

variability

The amount of dispersion for scores around some central value.

variable

Any event, situation, behaviour, or individual characteristic that varies —that is, can differ in some way (e.g., quantity, size, degree).

variance

A measure of the variability of scores about a mean. The variance is calculated by taking the difference between each score and the group mean, squaring these differences, and dividing the sum of these squares by the number of scores.

Web of Science

A database that also allows for cited reference searches, finding articles that cite a particular article.

“yea-saying” or “nay-saying” response sets

The tendency for some survey respondents to agree (yea) or disagree (nay) with the vast majority of questions being asked, regardless of the question content. This introduces error into the measure.

generalization

The ability for a finding based upon a sample of participants to tell us about the wider population from which that sample was drawn. A key question of generalization is not only whether a finding generalizes, but to what populations it can be generalized.

meta-analysis

A method of aggregating the results of many past studies to see whether an overall effect is observed in the past literature.

empirical question

A question that can be answered through empiricism, or systematic observation.

ruling out alternative explanations

One of the criteria for making an appropriate causal inference; ensuring that there are no other explanations for what might have caused an outcome.

secondary use of data

Analyzing data collected for other purposes, separate from the current research aim.

split-half reliability

An assessment of internal consistency for a scale achieved by splitting a measure into two halves, then correlating performance on one half with performance on the other half.

mundane realism

The extent to which the experiences in a study resemble closely an experience of everyday life.

experimental realism

The extent to which the experiences in a study are experienced by participants as impactful and engaging.

normal distribution

A prevalent distribution of scores for continuous variables, in which the majority of scores cluster around the mean (or average), with fewer and fewer scores observed the further they fall from the mean.

coefficient omega

Another estimate of internal reliability for a measure, shown to be superior to Cronbach's alpha, that is growing in popularity.

inattentive responding

Responding to survey questions without reading the item content, providing answers that are not thoughtful responses to the questions asked.

Remarks

*Use the top significance level when you have predicted a specific directional difference (a one-tailed test; e.g., Group 1 will be greater than Group 2). Use the bottom significance level when you have predicted only that Group 1 will differ from Group 2 without specifying the direction of the difference (a two-tailed test).

*Divide the significance level in half if using a one-tailed test.

New annotation

The running head is a shortened version of the title (up to 50 characters, including spaces). The title page shows the words “Running head:” followed by a shortened version of the title in CAPS.

All pages are numbered consecutively, starting with the title page. Page numbers are flush right, on the same line as the running head. Running head is flush left and page number is flush right in header area of the page.

The title is usually no more than 12 words.

Janet Polivy

Title, author(s), and affiliation are centred and appear in the upper half of the page.

author-note

The author note, if required, would begin below the authors’ affiliations. Student papers usually do not require an author note.

The running head identified on the title page should be carried forward on every subsequent page. It should remain flush left in all uppercase letters. The words “Running head” are deleted on subsequent pages.

abstract

The section of a research report that summarizes the entire study.

There is no paragraph indentation in the abstract.

Punctuation such as periods and commas appear inside the quotes. Punctuation such as question marks, exclamation marks, colons, and semicolons that are not part of the quote appear outside the quotes.

Standard form: Citation of two authors, parenthetically (Author & Author, Year).

Standard form: Citation of three to five authors, first citation, parenthetically (Author, Author, & Author, Year).

When several references are cited together, alphabetize, and separate with a semicolon.

three or more authors

Standard form: Citation of three or more authors, after first citation in text, parenthetically (Author et al., Year).

Standard form: Citation of a single author, parenthetically (Author, Year).

Standard form for citation of two authors, in text: Name and Name (Year).

Method

The Method section begins immediately after the Introduction (no new page). The word “Method” is centred and boldface.

Subsection headings such as “Participants” are **boldface**, typed flush to the left margin, and set alone on the line.

Use numerals to express numbers 10 and above.

Use numerals to express numbers that are immediately followed by a unit of measurement (in this case, hr for hour). Also abbreviate min (minute), s (second), ms (millisecond). Do not abbreviate day, week, month, year.

This section describes the measures and materials used to conduct the study. They make up a sort of “ingredients” list for the study.

This additional level of subheadings begins the paragraph. They are **boldface**, indented, and separated from the text of the paragraph with a period.

Define unfamiliar acronyms when they are first used.

The anchors of scales (e.g., very slightly or not at all) are italicized.

This section describes exactly how the study was conducted.

Numbers less than 10 are expressed as words.

Numbers that are immediately followed by a unit of measurement are expressed as numerals, as are numbers that represent time.

results

The section of a research report in which the researcher presents the findings.

When the outcome of a statistical test is presented, the name of the test is italicized and followed by the degrees of freedom in parentheses. The *p* refers to the probability of obtaining the results if the null hypothesis is correct.

Generally, exact probabilities are shown except when $p < .001$. Sometimes it is appropriate to use “ns” to indicate that a result is non-significant.

Statistical symbols

Statistical symbols, such as *M* for the mean, are italicized.

Greek letter

Greek letter *eta*, an indicator of effect size (squared value).

Any figures or tables must be mentioned in the text.

discussion

The section of a research report in which the researcher considers the research results from various perspectives.

Quotation marks are placed after the period.

APA strongly encourages writers to use past tense when reporting procedures and results.

references

A list of all sources cited in the body of a paper, sorted alphabetically by the last name of the first author of the cited material, formatted in APA style and including all information required to locate each source. Forms the last section of the APA format manuscript style.

A comma always follows the first author's initial, even if only two authors are listed.

doi

APA style recommends including the DOI (digital object identifier) if it is available.

APA recommends using the en-dash symbol between page numbers rather than a simple hyphen. The Microsoft Word shortcut to insert the en dash is Control-hyphen. This is probably not necessary for student papers.

Each reference begins on a new line and is considered a paragraph. The paragraph is a hanging indent, in which the first line is flush to the left margin and subsequent lines are indented.

Titles of books and journals are italicized, as are the volume numbers of journals.

&

Note that “&” (the ampersand symbol) is used for multiple authors throughout the References section.

Note capitalization: First word and first word following a colon, plus any proper nouns.

When the same set of authors is included multiple times, the entries are ordered by date, from oldest to newest.

The first line of the page should include the table number.

The next double-spaced line should include the table title, which should be italicized, with all major words capitalized. No period is required.

Only the first word of headings within the table are capitalized. Sections of the table are separated by horizontal lines. Do not use vertical lines.

A horizontal line should separate column headers from data presented in the table.

Include another horizontal line below the last row of information.

A note below the table is optional. The note may provide additional information such as an explanation of abbreviations or specific group differences.

Figure caption. Note italics for figure number. Caption may include more than one sentence.