

RESEARCH
IN THE CLOUD:
An Introduction to Modern
Methods in Behavioral
Science



CAMBRIDGE
UNIVERSITY PRESS

Aaron J. Moss, PhD
Leib Litman, PhD
Jonathan Robinson, PhD

Research in the Cloud: An Introduction to Modern Methods in Behavioral Science

A Project-Oriented Approach

Aaron J. Moss, Ph.D.^{1,2}, Jonathan Robinson, Ph.D.^{2,3}, & Leib Litman, Ph.D.^{2,3}

¹Siena College, ²CloudResearch, ³Lander College

Preview

Brief Contents

<i>Table of Contents</i>	<i>page</i> iii
<i>List of Figures</i>	xvi
<i>List of Tables</i>	xxv
<i>Preface</i>	xxvi
<i>Author Information</i>	xxx
<i>Introduction</i>	xxxii
<u>Part I: Fundamentals of Behavioral Research</u>	1
1: The Study of People	2
2: Behavioral Science Tools	22
3: Descriptive Research	56
4: Measurement	93
5: Correlational Research	129
6: Causal Inference	167
7: Experimental Research	209
8: AI Tools for Qualitative and Quantitative Research	257
<u>Part II: Conducting Online Research</u>	285
9: Understanding Online Participant Platforms	286
10: Data Quality	315
11: Data Quality Solutions	342
12: Data Cleaning	376
13: Best Practices for Designing Studies	396
14: Best Practices for Setting up and Launching Online Studies	428
15: Ethics and the Participant Experience	450
16: Writing	475

Table of Contents

<u>Part I: Fundamentals of Behavioral Research</u>	1
Chapter 1 – The Study of People	2
Module 1.1: Doing Research, Not Just Reading About It	4
Research Activity 1.1: Your First Research Project	4
Taking a Personality Test	6
Drawing Connections to Behavioral Research	7
Research Activity 1.2: Your First Data Analysis: Visualizing Personality Scores	9
Module 1.2: From Curiosity to Theory: Building Scientific Knowledge	12
The Making of a Scientific Theory: The Big Five Theory of Personality	12
What Makes a Theory Good?	13
Theory vs. Hypothesis	14
Testing Hypotheses, Refining Theories	15
Module 1.3: Behavioral Science in the Real-World	18
What is Behavioral Science?	18
The Disciplines of Behavioral Science	19
Where Behavioral Science Happens: Labs, Businesses, Governments, and Beyond	20
Summary	21
Chapter 2 – Behavioral Science Tools	22
Module 2.1: Tools for Finding Research Participants	25
Introduction to Online Research	25
The Role of Participants in Science	25
Research Activity 2.1: Participating in Online Studies	26
From Participant to Researcher	32
The Research Cycle	33

Module 2.2: Tools for Literature Review	35
The Value of a Literature Review	35
Research Activity 2.2: Conducting a Literature Review	36
Connecting to Peer-Reviewed Research	38
The Pyramid of Knowledge	39
Module 2.3: Tools for Creating Studies	41
Survey Platforms: Gathering Self-Reported Data	41
Performance Tools: Measuring Cognitive Processes	44
Physiological Measurement	45
Behavioral Measurement: Capturing Real-World Interactions	47
Module 2.4: Tools for Analyzing Data	49
Traditional Tools for Data Analysis	49
AI-Powered Tools for Data Analysis	51
Module 2.5: Tools for Sharing Research	52
Summary	53
Chapter 3 – Descriptive Research	55
Module 3.1: Basics of Measurement	58
What are Variables?	58
Conceptual vs. Operational Definitions	59
Three Types of Research: Descriptive, Correlational, Experimental	61
Module 3.2: The Power of Description	64
What is Descriptive Research?	64
A Real-World Example: Describing Mental Health Among U.S. Adults	64
Gathering Descriptive Data with Behavioral Sensors	65

Module 3.3: Designing a Descriptive Survey	67
Describing Reactions to a Moral Dilemma	67
Choosing What to Measure	69
Research Activity 3.1: Creating a Study in Qualtrics: the Heinz Dilemma	69
Gathering Descriptive Data Matched to the U.S. Census	79
Module 3.4: Analyzing Descriptive Data	80
Downloading the Heinz Dilemma Data from OSF	80
Research Activity 3.2: Analyzing People’s Responses to the Heinz Dilemma	81
What Descriptive Data Reveal about Moral Judgments	83
Module 3.5: Conducting Your Own Descriptive Project	85
Research Activity 3.3: Creating Your own Descriptive Study (and putting it in Qualtrics)	85
Gathering Descriptive Data Online	86
Summary	90
Chapter 4 – Measurement	92
Module 4.1: Measurement in the Behavioral Sciences	95
Research Activity 4.1: Working with Measurement Instruments	96
Calculating Total Scores: Anxiety, Depression, Trauma, and Sleep Disturbance	97
Visualizing the Distributions of Anxiety, Depression and Trauma for an Online Sample of 500 People	99
What the Measures Reveal	100
Module 4.2: Finding and Creating Measurement Scales	101
Research Activity 4.2: Finding Existing Measures	102
Reporting on the Measure You Found	102

Creating New Measures with AI	102
Strategies for Writing Strong Scale Items	106
Research Activity 4.3: Designing Your Own Measure	106
Using AI to Generate Scale Items	107
Evaluating Scale Items with AI	109
Assembling Your Scale	110
Reporting on Your Work	110
Module 4.3: Reliability and Validity	112
An Example: Establishing Reliability and Validity of the GAD-7 Anxiety Scale	113
Types of Reliability	114
Types of Validity	118
Module 4.4: Scales of Measurement	122
Nominal Measurement	123
Ordinal Measurement	123
Interval Measurement	123
Ratio Measurement	124
Why These Scale Types Matter in Research	124
Research Activity 4.4: Classify and Work with Demographic Variables in SPSS	126
Summary	127
Chapter 5 – Correlational Research	129
Module 5.1: What Do Correlations Tell Us?	132
Understanding Correlations	132
Research Activity 5.1: Measuring a Positive Correlation	133
Anxiety and Depression in the Real-World	134
Research Activity 5.2: Measuring a Negative Correlation	138
Age and Emotional Distress	139

When a Correlation Counts: Understanding Statistical Significance	139
Research Activity 5.3: Examining a Correlation Matrix	142
Module 5.2: Different Types of Associations	146
Associations between Categorical and Continuous Variables	147
Research Activity 5.4: Comparing Gender Differences in Depression	148
Conducting a <i>t</i> -test	149
Making Sense of Effect Sizes	150
Associations between Two Categorical Variables	151
Research Activity 5.5: Comparing Categorical Variables	152
Module 5.3: Guided Research Project: Morality and the Heinz Dilemma	155
What are We Studying: Project Goals and Big Questions	155
Part 1: Frame Your Hypotheses	156
Part 2: Design, Materials, and Methods	157
Part 3: Analyze the Data	159
Part 4: Interpret the Findings	161
Part 5: Report the Relationship Between Moral Foundations and People's Decisions	161
Module 5.4: Designing Your Own Correlational Study	163
Step 1: Craft Your Question and Study Design	164
Step 2: Collect Data	164
Step 3: Analyzing Your Results	164
Step 4: Interpret and Share What You Found	164
Summary	165
Chapter 6 – Causal Inference	167
Module 6.1: Controlling for Third Variables	170
Why Control Matters: Strengthening Internal Validity	172
A Thought Experiment: How to Control for Third Variables	173

Controlling Third Variables with ANCOVA: Marital Status, Depression, and Age	173
Activity 6.1 – Statistically Controlling Third Variables	178
Describing the Relationship between Marital Status and Depression after	
Controlling for Age	182
Controlling for Third Variables with Regression	182
Partial Correlations Explained	184
Research Activity 6.2: Multiple Regression in Action	185
Reporting Results: What Does the Regression Show?	187
Controlling for Multiple Third Variables at Once	188
Research Activity 6.3: Anxiety and Depression, Controlling for Trauma, Sleep,	
and More	189
Reporting What these Controls Reveal about Anxiety and Depression	190
Why Statistical Control Strengthens Causal Claims	191
Module 6.2: The Directionality of Cause and Effect	193
The Directionality Problem: Which Comes First?	193
Why Temporal Precedence Matters	194
Establishing Temporal Precedence with Longitudinal Research	195
Research Activity 6.4: Correlation between Anxiety and Depression one Year	
Later: Time-Lagged Correlations	198
Combining Approaches: Multiple Controls in Longitudinal Research	200
Module 6.3: Designing Your Own Causal Inference Study	203
From Correlation to Causal Inference: Your Research Project	203
Designing and Conducting a Causal Inference Study	203
Writing it Up: Telling the Story of Your Study	206
Summary	207
Chapter 7 – Experimental Research	209

Module 7.1: How Experiments Establish Causality	212
What is an Experiment?	212
Key Elements of the Experimental Method	212
The Logic of Random Assignment	214
Sample Size and Replication	216
Experiments vs. Correlational Studies	216
Where Experiments Appear in the Real-World	217
Research Activity 7.1: Design an Experiment	218
Creating Variations of the God Views Experiment	219
Module 7.2: Guided Project: Can Perspective-Taking Shift Moral Judgment?	221
Part 1: Understanding Perspective-Taking?	221
Part 2: Research Design, Materials, and Methods	223
Part 3: Data Collection	225
Part 4: Analyzing What You Found	226
Portfolio Entry: Report a Two-Group Randomized Experiment	229
Your Turn: Exploring the Effects of Perspective-Taking	229
Module 7.3: Variations on Experimental Design: Repeated Measures Experiments	231
Guided Research Project: How Consequences Shape Moral Judgments	233
Part 1: Understanding the Repeated Measures Question	234
Part 2: Research Design, Materials, and Methods	234
Part 3: Data Collection	236
Part 4: Analyzing What You Found	237
Portfolio Entry: Reporting the Results of Different Drug Outcomes	242
Module 7.4: Beyond Simple Experiments: Factorial Designs	243
Guided Research Project: When Does Perspective-Taking Work?	245
Part 1: Understanding the Factorial Research Question	246
Part 2: Research Design, Materials, and Methods	246
Part 3: Data Collection	249

Part 4: Analyzing What You Found	249
Your Turn: Exploring Moderators of Perspective-Taking	254
Summary	255
Chapter 8 – AI-Survey Tools for Qualitative and Quantitative Research	257
Module 8.1: The AI Revolution in Survey Research: From Static Surveys to Dynamic Conversations	260
Qualitative, Mixed Methods, and Quantitative Research	266
The Role of AI in Qualitative and Mixed Methods Research	267
Module 8.2: AI and Mixed Methods Research: A Study of Moral Reasoning Using the Heinz Dilemma	269
AI-Conducted Interviews: A Conversation about Morality in the Heinz Dilemma	272
AI-Based Approaches to Developing a Coding System	273
AI Tools for Mixed Methods Analyses	276
Using AI Tools for Thematic Analyses and Discovering Unexpected Patterns	277
The Researcher’s Role in AI-Driven Research	278
Summary	278
<u>Part II: Conducting Online Research</u>	285
Chapter 9 – Finding Participants Online	286
Module 9.1: Options for Online Participant Recruitment	289
A Brief History of Finding Participants	289
Market Research Panels	292
Researcher-Centric Online Platforms	300
Module 9.2: Representativeness in Online Sampling	306
What is a Representative Sample?	306

Two Approaches to Sampling: Probability and Non-Probability	307
Probability Sampling	307
Non-Probability Sampling	308
Why Behavioral Research Often Relies on Non-Probability Samples	309
The Fit-for-Purpose Framework: Matching Sample to Research Goals	311
Summary	313
Chapter 10 – Data Quality	315
Module 10.1: Data Quality in Online Research	319
Yea-Saying—A Common Behavior in Online Fraud	322
The Global Network of Research Fraud	323
The Rise and Fall of Mechanical Turk	324
High Quality Online Platforms	327
Module 10.2: How Data Quality Affects Research	329
How Bad Data Affect Descriptive Research	329
How Bad Data Affect Correlational Research	330
How Bad Data Affect Experimental Research	335
Summary	338
Research Activity 10.1: Data Quality in the Wild	339
Chapter 11 – Data Quality Solutions	342
Module 11.1: Detecting Fraudulent Responses	344
Detecting Yea-Saying	344
Using Benchmarks to Validate Attention Checks	347
Choosing an Exclusion Threshold	351
Open-Ended Validation	356

Module 11.2: Types of Attention Checks	358
What Are Attention Check Questions?	358
Types of Attention Checks	358
Issues With Attention Checks	365
A Strategy for Implementing Attention Checks	367
 Module 11.3: Implementing Data Quality Measures in Your Research	 368
Step 1: What Kind of Study Are You Conducting?	369
Step 2: Where will you Find Participants?	370
Step 3: How Many Checks Will You Use?	371
Step 4: What Criteria Will You Use to Exclude People?	372
Step 5: Will you Include any Benchmarking items or Open Ends?	373
 Summary	 373
 Chapter 12 – Data Cleaning	 376
Module 12.1: The Fundamental of Data Screening	379
The Purpose of Data Screening	379
Removing Versus Rejecting Participants	379
Evaluating Attention Checks	380
Computing an Attention Check Score	381
Evaluating Open-Ended Responses	385
Excluding Participants	386
 Module 12.2: Advanced Techniques of Data Screening	 388
What Advanced Methods Do	388
Why Advanced Methods are Hard to Implement	389
When Advanced Methods Might be Worth Considering	390
A Final Word: Avoiding False Positives	391
 Module 12.3: Guided Application: Cleaning Data from a Product Development Study	 392

Study Background: Fire Eaters, Spicy Snacks, and Brand Reputation	392
Assignment: Clean the Kellogg's Survey Data	393
Summary	394
Chapter 13 – Best Practices for Designing a Study	396
Module 13.1: Understanding Survey Platforms and Their Features	399
What is a Survey Platform?	399
Finding a Survey Platform	400
Survey Elements	400
Organizing Elements within Blocks	408
Branching and Skip Logic	411
Module 13.2: Best Practices for Survey Design	413
Matrices or Standalone Questions?	414
How Many Response Options?	417
Direction of Response Options	419
Vertical or Horizontal Answer Scales?	420
Open or Closed-Ended Items?	421
Other Survey Options	422
Summary	427
Chapter 14 – Best Practices for Setting up and Launching Online Studies	428
Module 14.1: Setting up and Launching an Online Study	430
Best Practices for Setting Up an Online Project	430
How to Name a Study	431
How and Where to Describe Your Project	432
How to Anonymously Follow Up With Participants	433
How to Figure Out Fair Payments	436

How to Find the Participants You Need	436
How to Increase Representativeness	439
When is the Best Time to Launch a Study	440
How and When to Communicate with Participants	441
Best Practices After a Project is Launched	441
Conducting Successful Longitudinal Studies	445
Summary	448
Chapter 15 – Ethics and the Participant Experience	450
Module 15.1: The Basis of Ethical Research	453
Risk of Harm in Online Research Studies	455
Research on Sensitive Topics Online	457
Fair Payment and Rejecting Participants	458
Protecting Privacy	460
Other General Best Practices for Online Studies	461
Module 15.2: Conducting Ethical Research: Writing an IRB Application	464
The Purpose of an IRB Review	464
Common Components of an IRB Proposal: An Example Study on Eating, Sleep, and Depression	464
Research Design and Methodology	465
Subject Population	466
Recruitment Methods	466
Informed Consent Process	468
Risks and Protections	468
Protection and Risk Mitigation	469
Benefits	470
Compensation/Payment	471
Example Consent Form	471

Summary	473
Chapter 16 – Writing	475
Module 16.1: Writing A Method Section	479
The Purpose of the 'Method' Section	479
Parts of a Method Section	479
Participants and Design	480
Procedure	482
Apparatus	485
Measures and Outcomes	485
Reporting Data Cleaning	487
Open Science, Data Sharing, and Transparency	488
Example Method Section	490
Module 16.2: Writing Advice	493
The Transaction	493
The Opening	493
Imitation	495
Actions and Ideas	495
Jargon	496
Ideas First, (Scientists Second)	497
Punctuation	498
Ending Strong	498
Recommended Books	499

List of Figures

Figure 1.1. Average scores from the Ten Item Personality Inventory, TIPI.

Figure 1.2. Percentile rank scores from the TIPI.

Figure 1.3. Measurement transforms theoretical constructs into data that can be analyzed.

Figure 1.4. A spreadsheet of student data.

Figure 1.5. Elements of a good theory.

Figure 1.6. A depiction of how theory forms through the scientific process.

Figure 2.1. The behavioral science toolkit.

Figure 2.2. The welcome message on Connect.

Figure 2.3. Demographic profile questions on Connect.

Figure 2.4. The participant dashboard on Connect.

Figure 2.5. The project preview page.

Figure 2.6. Logging into the researcher side of Connect.

Figure 2.7. Click "Create Project" to set-up a new study.

Figure 2.8. Search results on Google Scholar.

Figure 2.9. The pyramid of scientific knowledge.

Figure 2.10. Tools for creating four different types of studies.

Figure 2.11. A matrix question measuring the Big Five personality traits in Qualtrics.

Figure 2.12. How matrix questions appear to participants.

Figure 2.13. Tools that measure cognitive processes.

Figure 2.14. An electroencephalograph is an example of physiological measurement.

Figure 2.15. Measurement of real-world behavioral data from Facebook.

Figure 2.16. The data view in SPSS.

Figure 2.17. The Open Science Framework supports the research process.

Figure 3.1. Left panel: Marathon finishing times. Right panel: A 'bell curve.'

Figure 3.2. The prevalence of mental illness among U.S. adults in 2021.

Figure 3.3. Data on shower duration.

Figure 3.4. A view of the Survey Builder in Qualtrics.

Figure 3.5. A picture of blocks within the survey builder.

Figure 3.6. The welcome page of a survey.

Figure 3.7. The Heinz dilemma within the survey builder.

Figure 3.8. The Heinz dilemma as it appears to participants.

Figure 3.9. A question about whether Heinz should have stolen the drug.

Figure 3.10. A question asking people how acceptable it was for Heinz to steal the drug.

Figure 3.11. Three steps to adding demographic questions in Qualtrics.

Figure 3.12. An "anonymous survey link."

Figure 3.13. The "Files" tab will take you to folders with files you can download for each project.

Figure 3.14. A folder with files to download from the Open Science Framework.

Figure 3.15. Results of the descriptive study.

Figure 3.16. Click "Create Project" to start working on a new study.

Figure 3.17. Basic information for creating a project.

Figure 3.18. The space to paste a project link.

Figure 3.19. Options for sampling participants on Connect.

Figure 3.20. Applying the Census Matched Template.

Figure 3.21. The redirect URL provided by Connect

Figure 3.22. Configuring the redirect in Qualtrics.

Figure 4.1. Anxiety scores in SPSS.

Figure 4.2. Distribution of anxiety scores.

Figure 4.3. The relationship between reliability and validity.

Figure 4.4. Types of reliability.

Figure 4.5. Illustration of internal consistency.

Figure 5.1. A positive correlation in a scatterplot.

Figure 5.2. Scatterplot between anxiety and depression.

Figure 5.3. A negative correlation in a scatterplot.

Figure 5.4. A scatterplot depicting no relationship between two variables.

Figure 5.5. A correlation matrix.

Figure 5.6. Three types of association.

Figure 5.7. Differences in depression scores between men and women.

Figure 6.1. The directionality problem with violent video games and violent behavior.

Figure 6.2. A third variable may explain why violent video games and violent behavior are correlated.

Figure 6.3. Association between marital status and depression. Married people are less depressed than single people.

Figure 6.4. A demonstration of the third-variable problem. Age may explain why married people are less depressed than those who are single.

Figure 6.5. Controlling for age involves making age into a constant and examining whether the association between anxiety and depression remains intact.

Figure 6.6. Marital status is no longer a statistically significant predictor of depression when age is statistically made into a constant.

Figure 6.7. A third variable such as trauma may explain why anxiety and depression are correlated.

Figure 6.8. The relationship between anxiety and depression remains statistically significant even after statistically making trauma into a constant.

Figure 6.9. A statistical model examining the correlation between anxiety and depression while holding trauma, sleep, income, and education constant.

Figure 6.10. The relationship between anxiety and depression remains statistically significant even after statistically making trauma, sleep, income, and education into constants.

Figure 6.11. An illustration of the directionality problem.

Figure 6.12. Longitudinal designs yield a variety of correlations.

Figure 6.13. An example of a cross-lag correlation: depression measured at Time 1 predicts anxiety one year later.

Figure 6.14. Cross-lag correlation between depression at Time 1 and anxiety one year later, while holding anxiety constant at Time 1.

Figure 6.15. Cross-lag correlation between Anxiety at Time 1 and Depression one year later, having controlled for initial levels of anxiety.

Figure 6.16. Predicting anxiety at Time 2 from depression at Time 1 while controlling for multiple variables at Time 1.

Figure 7.1. Experimental design of the God views study.

Figure 7.2. Random assignment distributes all variables equally across conditions. In effect, this makes all variables constants across conditions – the means of variables are equal across conditions.

Figure 7.3. Randomization of participants to blocks within the Qualtrics 'Survey flow.'

Figure 7.4. The design of the perspective-taking experiment.

Figure 7.5. Randomization within the perspective-taking experiment.

Figure 7.6. Differences between independent groups (between-subjects) and repeated measures (within-subjects) experiments.

Figure 7.7. The three conditions are presented in a randomly determined order for each participant.

Figure 7.8. The table of pairwise comparisons shows which condition means are significantly different from one another.

Figure 7.9. Results of the within subjects experimental design.

Figure 7.10. Results of a factorial study showing a main effect of religiosity and an interaction between religiosity and the God perspective manipulation.

Figure 7.11. A 2×2 factorial design yields four experimental conditions

Figure 7.12. ANOVA results for a factorial design.

Figure 7.13. Pairwise comparisons test the effect of one independent variable within the level of another.

Figure 7.14. The interaction between wealth and perspective-taking in the Heinz dilemma.

Figure 8.1. A traditional fixed-response question about reasons for drinking alcohol

Figure 8.2. A hypothetical conversation with a respondent exploring drinking behavior through open-ended conversation

Figure 8.3. Motivational themes for drinking alcohol extracted by AI from conversational interviews

Figure 8.4. Diagram comparing quantitative and qualitative approaches, highlighting advantages of qualitative methods

Figure 8.5. Engage platform following close-ended with open-ended questions

Figure 8.6. Sample interview between an AI interviewer and a human participant about the Heinz dilemma

Figure 8.7. Instructions provided to the AI survey system for analyzing qualitative data

Figure 8.8. Example of a participant's response categorized as Stage 2 of moral reasoning with AI rationale

Figure 8.9. Example of a participant's response categorized as Stage 4 of moral reasoning with AI rationale

Figure 8.10. AI tagging frequency of different moral reasoning themes

Figure 9.1. Milgram's newspaper ad from New Haven Register

Figure 9.2. Annual online survey completion statistics

Figure 9.3. Three tiers of the market research ecosystem

Figure 9.4. How participants are directed to studies within market research ecosystem

Figure 9.5. Connect platform compensation setting interface

Figure 9.6. Connect Conversations center interface

Figure 10.1. Joe Rogan and Neil deGrasse Tyson discussing false research findings

Figure 10.2. A look inside a survey click farm

Figure 10.3. Reddit post explaining how to sell participant credentials to fraudsters

Figure 10.4. Facebook post seeking to rent a participant ID on MTurk

Figure 10.5. Dangerous cleaning behaviors by participant quality

Figure 10.6. Correlation between education and social anxiety by participant quality

Figure 10.7. Facebook use and depression correlation by participant quality

Figure 10.8. Trolley dilemma responses by participant group

Figure 10.9. Soda task results across three data quality conditions

Figure 11.1. Proportion of problematic and fraudulent respondents across three online platforms

Figure 11.2. Two sets of questions: attention checks vs. benchmarking questions

Figure 11.3. Reports of scuba diving among three groups

Figure 11.4. Reported rates of engaging in at least 1 of 3 benchmarking activities

Figure 11.5. Benchmarking activity rates with more stringent threshold

Figure 11.6. Benchmarking activities with stringent threshold

Figure 11.7. How reported rate of rare behaviors changes as people fail more yea-saying questions

Figure 11.8. Matrix of multiple-choice questions with embedded instructed response item

Figure 11.9. Questions that can guide your plan to protect data quality

Figure 12.1. Export Data options in Qualtrics

Figure 12.2. SPSS Recode Variables dialog

Figure 12.3. SPSS Recode Values dialog

Figure 12.4. SPSS Compute Variable dialog

Figure 12.5. SPSS Frequency output table

Figure 13.1. Images as survey elements for product presentation

Figure 13.2. Ranking question for message effectiveness

Figure 13.3. Drag and drop question for message categorization

Figure 13.4. Heat map results showing participant interactions

Figure 13.5. Slider scale example for confidence assessment

Figure 13.6. Example block with multiple survey elements

Figure 13.7. Between-subjects vs. within-subjects randomization

Figure 13.8. Adding branches in survey flow

Figure 13.9. Question matrix with Big Five Personality items

Figure 13.10. GAD-7 questions in single matrix format

Figure 13.11. Response scales with and without middle options

Figure 13.12. Vertical vs. horizontal scale alignment

Figure 13.13. Embedded data capture interface

Figure 14.1. Connect dashboard showing navigation tabs on left and study listings in center

Figure 14.2. The text entered in the "project description" box appears on the participant dashboard as they browse projects

Figure 14.3. After choosing to view a project, participants are presented with special instructions from the researcher

Figure 14.4. Embedded data allows you to capture people's participant IDs without asking for them in the survey

Figure 14.5. The Connect interface allows researchers to select specific characteristics for recruiting participants.

Figure 14.6. Each bin or bucket within a quota can be adjusted to the desired percentage of the sample. In this case, the sample will be 33.3% Republican, 33.3% Democratic and 33.3% Independent

Figure 14.7. The "Census Matched Template" adds quotas matched to the latest US Census figures for age, race, ethnicity, and gender

Figure 14.8. The conversations center allows researchers and participants to communicate with each other while maintaining participant's anonymity

Figure 14.9. When a study is live, the project dashboard displays information that is useful for monitoring study progress

Figure 14.10. Reports of broken content or other technical errors from participants can be invaluable, sometimes preventing scarce resources from being wasted on a project that contains errors

Figure 15.1 - A depiction of how the core principles of the Belmont Report get applied in specific parts of the research process

Figure 15.2 - A sample study posting on the Connect platform

Figure 16.1 - Papers written in APA style follow a formula.

List of Tables

Table 1.1. The Disciplines of Behavioral Science

Table 3.1. Items from the GAD-7.

Table 4.1. Sample responses to the GAD-7.

Table 4.2. Common question types.

Table 4.3. AI-generated items to measure academic stress.

Table 4.4. Inter-item correlations.

Table 4.5. Cronbach's alpha thresholds.

Table 4.6. Types of measurement.

Table 5.1. Conventions for the strength of a correlation coefficient.

Table 5.2. Conventions for interpreting Cohen's d effect sizes.

Table 9.1. Panels that supply participants to market research aggregators

Table 9.2. The strengths and weaknesses of market research panels

Table 9.3. The strengths and weaknesses of crowdsourcing platforms for research

Table 10.2. Comparison of true U.S.-born adolescents, true immigrants, and those who falsely claimed immigrant status

Table 11.1. Five yea-saying questions from Reavey et al. (2024)

Table 15.1 - Examples of challenges participants face in online research

Table 16.1 - Demographic information from Robinson et al. (2019)

Preface

The most important thing I learned in almost thirty years of teaching experimental psychology is that the best way to learn, and the best way to teach, is by actively engaging in the research process. By “best” I don’t just mean effective in terms of achieving learning goals, but also in terms of being less boring and more fun for both students and teachers. This is just as true for advanced graduate students as it is for students in their first methods class—indeed on their very first day in that class. Research methods are best learned by doing research, and modern technology now makes it possible to provide this experience to students at all levels.

When I first started teaching in the 1990s, the music was better, but the opportunities for engaging students in real research were limited. Today, the landscape has fundamentally changed. With the advent of online participant recruitment, easy-to-use survey platforms, the Open Science Framework for sharing materials, and AI-assisted tools, it is possible to develop and launch a study, collect data from real participants, and download and analyze the results all within a single two-hour lab session. What once took an entire semester can now be accomplished in a typical class period.

The purpose of this book is to bring the most useful professional research tools into the classroom to create the best possible project-oriented learning experience for both students and professors. To do so, my co-authors and I have developed several methodological innovations that reimagine how research methods can be taught.

Structuring Classes as “CLABs”: Classroom + Lab Hybrids

The chapters in this book are structured around CLABs (Classroom-Laboratory hybrids), which are integrated sessions where theoretical concepts are immediately followed by hands-on projects. CLABs foster Collaborative Learning About Behavioral Science.

To facilitate project-oriented learning in CLABs, we have created ten research projects with real data collected from over 2,500 online participants specifically for this book. Part 1, Chapters 1-8, is geared toward the introductory student and combines traditional content with projects and assignments to bring the concepts to life and encourage independent learning. The content and projects are carefully scaffolded to support students towards incremental mastery of key

concepts. Part 2, Chapters 9-16, is geared towards advanced undergraduates, graduate students and researchers alike and can be used as a guide for conducting research online. It focuses on the participant experience, data quality, study design and ethical considerations in the online landscape. All materials, including Qualtrics files, surveys, datasets, and analysis instructions, are available through this book's Open Science Framework page. Every project is also accompanied by short, engaging videos that guide students through each step of the activity.

A Four-step Learning Progression

Each chapter follows a four-step learning progression: First, chapters introduce a theoretical concept (e.g. correlation). Second, students work with data to see real-world applications of that concept (e.g. conduct a correlation analysis between anxiety and depression in a 500-participant online sample and visualize the scatterplot). Third, students engage in a guided research project. As part of the guided project, they examine an already existing research study where the data have been collected, and the materials are available to work with. Students engage with these guided projects as a researcher would. They start by formulating a hypothesis prior to seeing the data, work with the Qualtrics file, download and analyze the data, examine whether the results support their hypotheses, and write a brief explanation of the findings. The final step is the creation of their own project, such as a correlational study created on Qualtrics. This four-step process builds skills incrementally and brings each methodological concept to life.

We have also incorporated the latest technological tools to facilitate students' ability to create their own projects. One of the most productive examples is the use of AI to create their own 7-10 item measurement scales while they learn about measurement. This approach, covered in Chapter 4, enables students to create their own measurement instruments to measure just about anything that interests them and allows them to explore and research a much wider set of topics than would otherwise be possible.

Finally, we have implemented a portfolio-based assessment approach where students build a cumulative record of their research accomplishments throughout the course. At the end of each small project, students copy and paste their SPSS output to their portfolio and then describe and interpret the findings. This creates a simple and effective way to assess participation and

engagement in each activity and also creates a repository of skills that the students acquire over the semester.

The result is a textbook that doesn't just tell students about research—it enables them to do research. In the first-class session, students dive in, with a simple and engaging data collection exercise. By mid-semester, they are typically designing and conducting original studies for which they create their own measurement scales and which they program on Qualtrics, collect the data online, and analyze using statistical software. And by the end of the course, they will have developed a comprehensive portfolio showcasing their ability to create their own measures, interpret the results of t-tests, multiple regression, and other statistical analyses, create randomized experiments on Qualtrics, formulate research questions, design studies, collect and analyze their own data, communicate research findings, and much more.

This approach reflects my deepest conviction about learning: we master what we practice, not what we read. The students I've taught over the years have consistently confirmed this principle. Those who thrive—whether they become researchers, practitioners, or professionals in entirely different fields—are those who have experienced research firsthand.

We feel this book adheres to a model of *transformational teaching*, (Slavich and Zimbardo, 2012), which involves dynamic relationships between students and teachers to facilitate not only learning but personal growth. In this approach, teachers function as “intellectual coaches” in which we provide a framework for learning and disseminate core concepts while allowing for independent and active learning. Mentoring and teaching should be transformational for both student and professor where both parties can learn from each other

Research in the Cloud brings that opportunity to every student. I invite you to join us in this reimagined approach to research methods education. Whether you are a student encountering these concepts for the first time or an instructor looking to transform your teaching, I hope this book helps you experience the excitement and insights that come from doing real research from day one.

Housekeeping for Instructors

A note about IRB's. Research that is conducted solely for educational purposes and is not intended for external dissemination, does not typically require IRB approval. Thus, students can generally conduct the projects outlined in this course without IRB approval. However, it is common for students to come up with creative and exciting projects that deserve to be showcased to a larger audience. In such cases, IRB approval will be warranted. Each instructor should speak with their institution's IRB to understand the requirements of their specific institution. Chapter 15 includes ethical considerations for online studies and best practices for writing consents and IRB proposals.

Leib Litman,

Professor of Psychology, Lander College

Preview

Author Information

Aaron Moss, PhD., is a social psychologist who studies online research methods. He has written over 20 peer-reviewed publications, and his work has been covered in The Wall Street Journal and The Harvard Business Review. He teaches at Siena College and lives in Albany, NY.

Dr. Jonathan Robinson, Computer Science Professor at Touro's Lander College and CloudResearch CTO, transforms complex technology into accessible tools. Author of "Conducting Online Research on Amazon Mechanical Turk and Beyond," he challenges students to surpass him while living his motto: "Dream and make it happen." His innovative work impacts research methodology.

Leib Litman, Ph.D., is Professor of Psychology at Lander College and Chief Research Officer at CloudResearch. He co-wrote *Conducting Online Research on Amazon Mechanical Turk and Beyond* (2020), and his research has been covered by the New York Times, the Wall Street Journal, Forbes, Harvard Business Review, and others.

Introduction

Leonardo da Vinci is famous for many things—his painting, drawing, and contributions to engineering, architecture, and scientific thought to name a few. Of all the noble pursuits he set his mind to, describing the tongue of a woodpecker is a puzzling one.

Walter Isaacson wrote a biography about Leonardo. After learning about the woodpecker in a notebook, Isaacson wondered: “Who on Earth would decide one day, for no apparent reason, that he wanted to know what the tongue of a woodpecker looks like?” The answer was Leonardo; his reason was curiosity.

Of all the motives behind scientific inquiry, curiosity may be the most virtuous. It’s what led Orville and Wilbur Wright—two bicycle mechanics who lacked scientific training—to conduct systematic tests that led to the first machine powered human flight. It’s what Einstein credited for his ability to grasp the fundamentals of the cosmos, writing, “I have no special talents, I am only passionately curious” (as quoted in Isaacson, 2008). Curiosity is, in fact, the motor behind many scientific advancements from Copernicus’s heliocentric view of the universe to Newton’s formulation of the laws of gravity, from Darwin’s theory of natural selection to Pasteur’s germ theory, and from Watson, Crick, and Franklin’s discovery of DNA’s structure to Jennifer Doudna’s development of CRISPR for editing that DNA. As a catalyst for discovery, curiosity is hard to beat.

While some people are curious about germs, genes, or galaxies, everyone is curious about the human mind. Questions about why people think, feel, and act as they do are central to everyday life and the focus of research in the behavioral sciences. Even a cursory glance at this branch of science reveals much to be curious about.

Why, for example, do people get bored? What leads to a meaningful life? How well do we know ourselves? What is the connection between the mind and the brain? Do humans have free will? What qualities do people want in a romantic partner? What causes mental illness? How well do attitudes predict behavior? Does religion promote morality? When do people feel regret? Does personality change with age? The interesting questions are endless.

And yet, discovering things about the human mind or behavior requires more than curiosity. Anyone who wants to know *why* people do what they do must also understand the methods behavioral scientists use to conduct research. These include things like how to form a research question, design a study, test a hypothesis, measure psychological concepts, identify patterns in data, demonstrate cause-and-effect relationships, and communicate research findings to others. In today's world, it also means understanding how research methods are integrated with the digital tools that have become essential to how behavioral science operates. Teaching the fundamentals of behavioral science and how to use its newest tools is the goal of this book.

How to Use This Book

This book was designed with a two-part structure to serve learners at different stages of their research journey. Part I explains the basics of behavioral research; Part II guides you through how to apply these methods online.

Part 1: Foundations of Behavioral Research

The first eight chapters provide a hands-on introduction to behavioral research methods. For students new to behavioral research, this is the starting point. We take a friendly, step-by-step approach to core concepts like forming research questions, understanding different types of studies, and collecting data. Throughout these chapters, students learn by doing—working through carefully designed activities that make abstract concepts concrete and manageable. More than just reading about research methods, students conduct small studies, analyze real data, and develop research skills. We created an entire video series to accompany these projects.

Part 2: Online Research Methods

The second half of the book provides a guide to conducting online research. This section can be used as a standalone guide for conducting online research. This section is valuable for students in an advanced methods courses, for people working on an honors thesis or independent research project, for graduate student's planning a master's thesis or doctoral dissertation, and for postdoctoral researcher or faculty members who have little experience with online research. The chapters in Part II describe the diverse ecosystem of online participant recruitment and look at topics like where to find online participants, how to manage data quality, best practices for setting up an online study, best practices for survey programming and data cleaning, and ethics.

The two-part structure of this book means different readers can approach the book in different ways. Students in an introductory methods course can focus on the research methods fundamentals covered in Part I while referencing Part II as necessary. Those who are specifically interested in online research methods, can start with Part II's applications to online research, referencing Part I where needed.

Regardless of starting point, this book will help you develop practical skills that are valued across academic and professional settings. You will learn to design studies, collect and analyze data, use modern research platforms, manage projects, and communicate findings effectively. These skills are increasingly sought after in fields ranging from market research and public policy to technology, law, and healthcare.

Just as Leonardo da Vinci's curiosity led him to make amazing discoveries about the world, this book will equip you with the tools to systematically investigate questions about the human mind. Whether you are pursuing an academic career or planning to apply these skills in industry, you will learn methods that professional researchers use every day in universities, businesses, non-profits, and government agencies.

Our goal is to help you develop not just technical knowledge, but the ability to satisfy your curiosity about human nature through rigorous research. As you work through this book, remember that each method you learn and each project you complete brings you closer to understanding the fascinating complexities of human thought and behavior. Have fun!

Part I: Fundamentals of Behavioral Research

Preview

Chapter 1: The Study of People

Trying to learn about human behavior simply by observing people in your life is like trying to understand the ocean by standing on the beach. You might notice the waves and the tide, see a few creatures on the shore, and feel the water's temperature, but most marine life will remain hidden from view. Similarly, casual observations of the people around you might reveal interesting behaviors and noteworthy patterns, but these observations only scratch the surface. They can't tell you *why* people do what they do or how reliable your observations are. To truly understand people, you need a systematic approach.

Unlike traditional research methods textbooks, this book will introduce you to behavioral research through hands-on activities. Instead of just reading about research methods, you will use the tools and methods researchers use when they conduct scientific investigations. This approach allows you to experience how behavioral scientists collect data, analyze results, and develop theories by participating in each step.

In Module 1.1, you will see how researchers transform abstract psychological characteristics into measurable variables. This will be your first activity and your first experience working with data. In Module 1.2, you will learn how researchers transform observations into theories, using the history of personality as an example. Finally, in Module 1.3, you will explore the diversity of the behavioral sciences, examining how different disciplines approach the study of human behavior. Through the activities in this chapter, you will discover how behavioral scientists use systematic methods to move beyond casual observations toward scientific explanations.

Be prepared: this chapter introduces you to a new way of thinking about people and a new approach to learning. The hands-on activities might feel different from traditional coursework, but they are designed to help you develop the knowledge and skills of a behavioral scientist. By completing each exercise, you will begin to understand the research process from the inside. You will also build knowledge and skills that will serve you well throughout this course and beyond.

Chapter 1 Outline: The Study of People

Module 1.1: Doing Research, Not Just Reading About It



Laboratory

Learn how to turn concepts into variables and work with data

- Take a short online personality test
- Learn what this test reveals about measuring psychological concepts
- Enter your data into a spreadsheet and see how scientists work with data

Module 1.2: From Curiosity to Theory: Building Scientific Knowledge



Reading

Explore how scientists develop theories to explain patterns in data

- Read about the development of personality theory
- Understand what makes a theory "good"
- Learn how scientists test hypotheses and refine theories

Module 1.3: Behavioral Science in the Real-World



Reading

Examine where behavioral science is practiced

- Explore what behavioral science is and where it is practiced
- Learn about the diverse fields of behavioral research

Doing Research, Not Just Reading About It

Module 1.1

Learn how to turn concepts into variables and work with data

Have you ever wondered why your roommate waits until the night before an exam to start studying while you begin days in advance? Or why some people thrive in large social gatherings whereas others feel drained? These aren't just idle curiosities—they are the kinds of questions that have launched entire fields of scientific research.

Traditionally, research methods textbooks begin by describing how behavioral scientists think about theories, how they weigh evidence, and the methods they use to understand people. But we are taking a different approach. We are going to let you experience what research is like.

In just a minute, you will participate in an exercise that assesses personality. While personality is a fascinating topic, our goal is to illustrate some fundamental ideas about behavioral science. These include how scientists measure psychological concepts, how they collect and organize data, and how they use that data to understand people's thoughts and behavior. After you complete the project, we will discuss the role of theories and evidence in behavioral science and we will describe the academic disciplines that follow this approach to research. By the end of the chapter, you will understand not only how behavioral research operates but some of the professions, disciplines, and settings where behavioral scientists study human behavior. Let's jump in!

Research Activity 1.1: Your First Research Project



If you read this book's Introduction—and, if you didn't our question is why not?—you know we want to teach you about research methods by having you do research. Here's your first chance.

We want you to take a short personality test known as the TIPI (pronounced tip-ee). TIPI stands for Ten Item Personality Inventory (Gosling et al., 2003), and it was designed to quickly measure The Big Five personality traits described in Box 1.1.

After you complete the test, you will receive a score for each dimension of your personality. Write these scores down or take a screenshot. You will use them later in the chapter.

Your scores on the TIPI will be the first example in this book of how behavioral scientists collect data to measure human characteristics. As you will see, turning abstract psychological characteristics like extraversion or conscientiousness into numbers is the first step in scientifically studying human behavior.

The Big Five Personality Traits

Openness to Experience: This trait captures a person's curiosity, creativity, and willingness to try new things. People high in openness tend to be imaginative, intellectual, and adventurous. They appreciate art and beauty, and enjoy exploring new ideas. People low in openness are typically more conventional, practical, and prefer routine.

Conscientiousness: This trait reflects a person's level of organization, dependability, and self-discipline. Highly conscientious people are typically goal-oriented, reliable, and methodical. They pay attention to detail and meet deadlines. People low in conscientiousness are more spontaneous and may struggle with organization and follow-through.

Extraversion: This trait describes a person's social energy and tendency to seek out interactions. Extraverts are outgoing, energetic, and gain energy from being around others. Introverts are more reserved, prefer smaller social interactions, and need alone time to recharge their energy.

Agreeableness: This trait reflects a person's interpersonal style, particularly compassion and empathy. Highly agreeable people are kind, sympathetic, and value harmony in relationships. They tend to be trusting and concerned with others' feelings. Those low in agreeableness are more competitive and may prioritize their own needs over others'.

Neuroticism (or Emotional Stability): This trait measures emotional reactivity and stability. People high in neuroticism experience more frequent negative emotions like anxiety, depression, and vulnerability to stress. Those low in neuroticism (high in emotional stability) are typically calm, resilient, and better at managing stress.

Box. 1.1. A description of the Big Five Personality traits.

Taking a Personality Test

To take the TIPI, go to: <https://psytests.org/big5/tipien.html>. Select the button that says “Begin,” and then on the next page select “Begin” again. Once the test starts, you will be asked ten questions. They should take less than a minute to complete.

Your results will look like Figure 1.1. Scores on each trait will range between 1 and 7. For example, the agreeableness score in Figure 1.1 is 6.5.

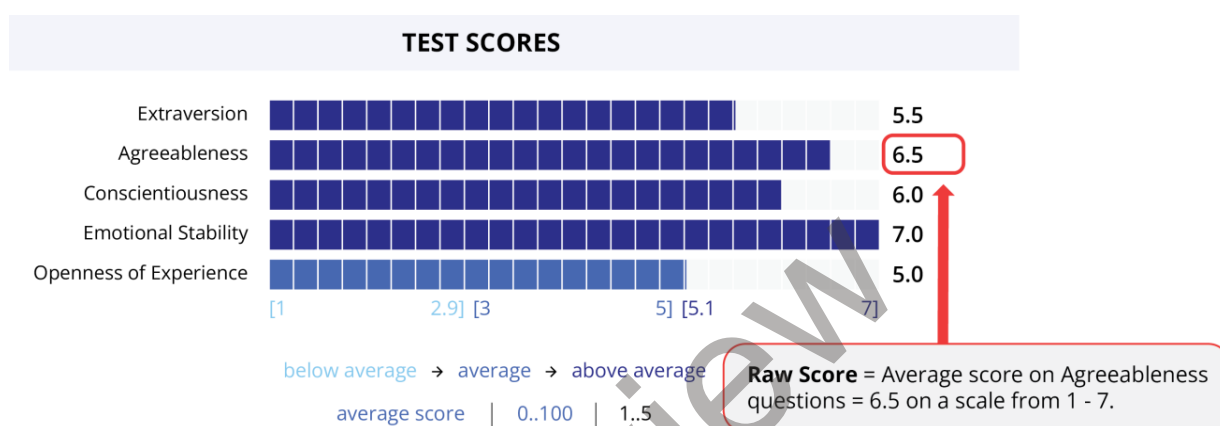


Figure 1.1. Average scores from the Ten Item Personality Inventory, TIPI.

You can toggle between raw scores and percentiles, as shown in Figure 1.2. A percentile represents the percentage of people whose scores fell below yours on each trait. Percentiles are interesting because they show where you stand relative to others. According to Figure 1.2, an agreeableness score of 6.5 is higher than 92% of people who have taken the test.

After reviewing your scores, remember to keep the results tab open, write down your average on each trait, or save a screenshot. You will need these soon.

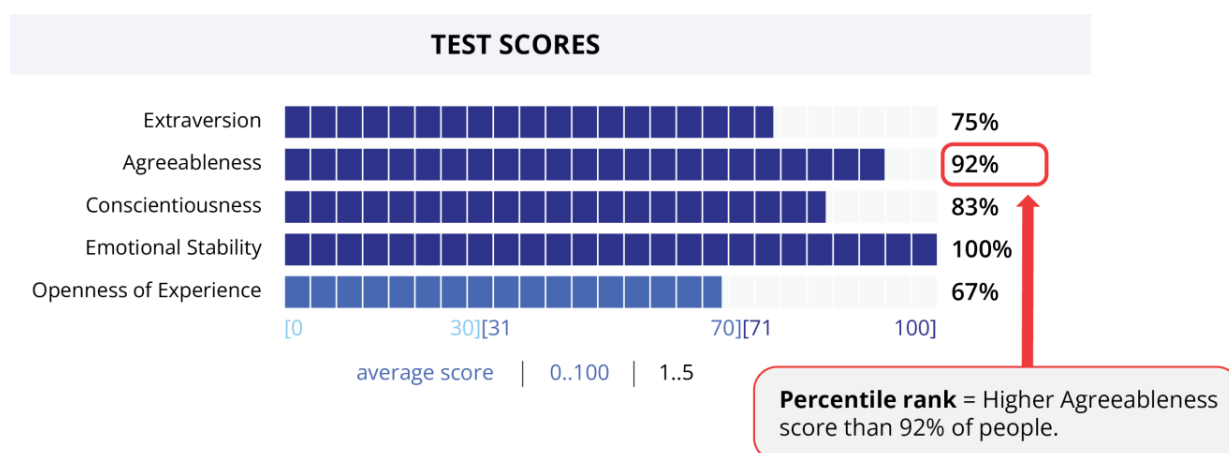


Figure 1.2. Percentile rank scores from the TIPI.

Drawing Connections to Behavioral Research

Let's explore what this exercise reveals about how behavioral scientists make measurements that are useful for scientific research (Figure 1.3).

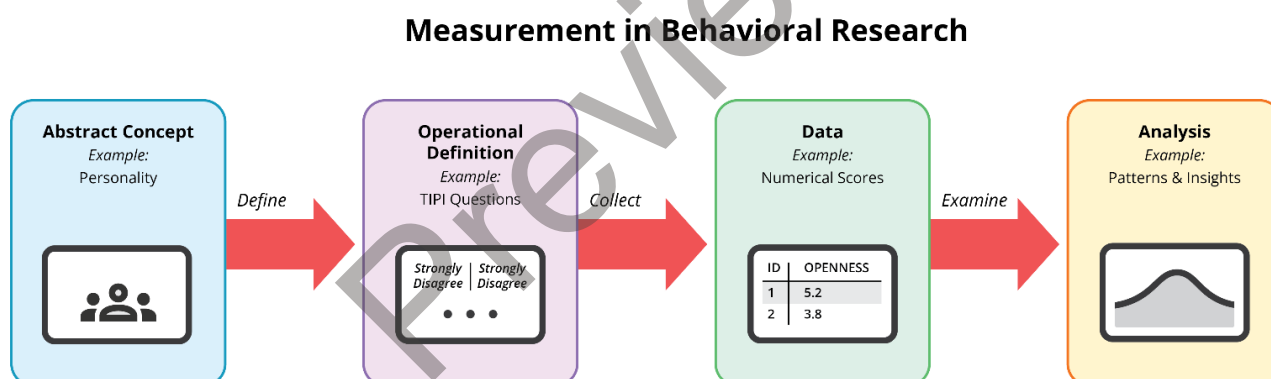


Figure 1.3. Measurement transforms theoretical constructs into data that can be analyzed.

First, behavioral research often requires measuring abstract characteristics about people. Unlike physical characteristics such as height, weight, time, or distance, there is no ruler or scale that measures a psychological construct like agreeableness. Instead, researchers must develop questionnaires and other measurement instruments. But developing a measure raises interesting questions, such as: what evidence is there this this particular measure—called an operational definition—is a good measure, and how does the researcher know if the measure is accurate? The answer to these questions involves evaluating **reliability** and **validity**, two concepts you will learn about in Chapter 4.

Second, measuring personality is an example of **quantitative research**. In quantitative research, scientists convert complex characteristics, like agreeableness or openness to experience, into numbers that can be analyzed. Each measured characteristic represents a **variable**, which is simply a measure on which people have different scores. To see **variability** in personality, all you need to do is compare your scores to other people. Are you higher or lower than others in extraversion, agreeableness, and emotional stability? Measuring variables makes it possible for behavioral scientists to systematically study how people differ and what those differences mean.

Third, once researchers can measure a characteristic, they can look for patterns in data. For instance, studies have found that people who score low in conscientiousness tend to receive more parking tickets, earn less money, and even have shorter lifespans than those who score high in conscientiousness (Alderotti et al., 2023; Bogg & Roberts, 2013). Conversely, people high in conscientiousness often experience greater career success, higher job satisfaction, and better physical health, including lower rates of chronic disease (Judge et al., 2013; Strickhouser et al., 2017; Wilmot & Ones, 2019). Research like this reveals how different variables are related, allowing scientists to discover fascinating patterns in human thought and behavior.

Finally, taking the TIPI has given you a glimpse at what it's like to be a research participant. Think about your experience: how seriously did you take the assessment? Did you have fun? Did the length of the test feel appropriate? Would your responses have been different if the test was longer? Your answers to these questions concern what is called participant engagement, which as every seasoned researcher knows, is as important to the quality of a study as good measures or sound study design. Throughout this book, we will explore how researchers balance scientific rigor with practical considerations like participant engagement.

Now that you understand some basics of behavioral research, try developing your own question about personality. For instance, which behaviors do you think extraversion or openness to experience might predict? What kinds of outcomes do you think are common among people low in emotional stability? Once you can form these kinds of questions, you will be a step closer to designing a study that can provide an answer.

Research Activity 1.2: Your First Data Analysis

Now that you have taken the TIPI, let's dig into some data.

For this exercise, we created a Google Sheet where you—and everyone else who reads this book—can enter your personality scores. You can access the sheet here: <https://bit.ly/4hojglZ>.

Once you open the sheet, you will see five columns (one for each trait) and several rows of student data (we occasionally clean up the file so it's easy to work with). The structure of this spreadsheet reflects how behavioral data is typically organized. No matter how complex the project is, researchers enter their data into a spreadsheet just like this, where each row represents a person and each column represents a variable. Once the data are ready, you can conduct statistical analyses and create figures to visualize the results.

To add your data to the thousands of students who have completed this exercise, enter your scores into each column, staying within a single row. When you enter a number in columns B through F, a new student ID will appear in column A. Figure 1.4 shows what the data should look like.

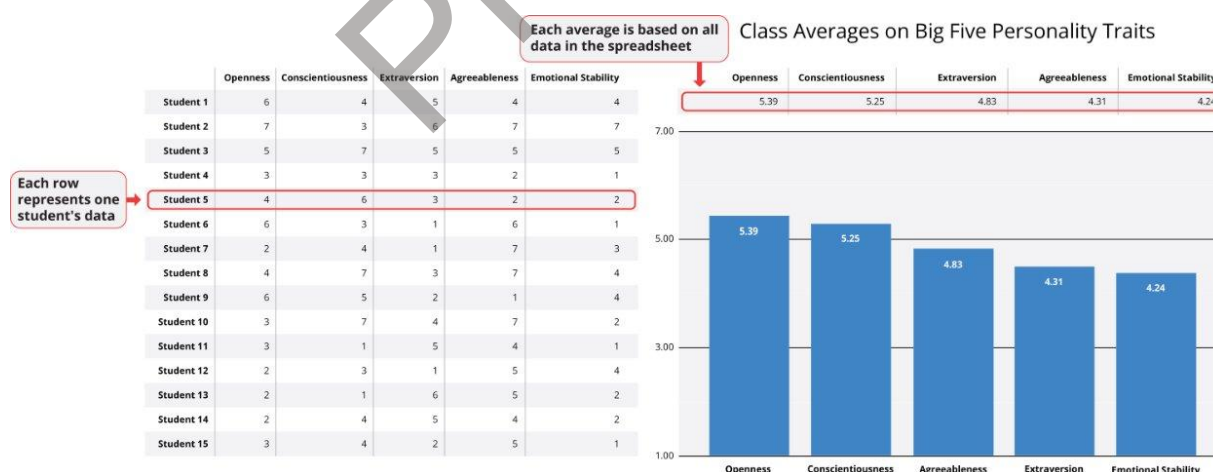


Figure 1.4. Left: TIPI scores organized in a spreadsheet. Each row represents one student's scores, and each column represents one of the Big Five personality traits. Right: A bar chart showing the average scores for each trait.

Row 2 in columns I through M contains a formula that computes the average for each trait (e.g., =AVERAGE(B:B)). When you add your data, the averages within each column and the bar chart will automatically update.

Even though your task is simple, you are working with data by entering numbers, computing averages, and examining the visual pattern across many people. Behavioral scientists do something similar in each study they conduct.

As you examine the data, consider these questions: Which personality trait has the highest average? Which has the lowest? Do these patterns match what you expected? How might the pattern change if the data represented the entire U.S. population rather than students enrolled in research methods? These are the kinds of questions researchers might ask as they analyze these data.

Later in this book, you will learn more sophisticated ways to collect, organize, and analyze data, but the basic processes will remain the same. Behavioral research always involves converting abstract constructs into numbers, collecting data from people, entering the data into a spreadsheet, analyzing the data, and creating visualizations to understand patterns and communicate about the implications for human behavior.



Stop and Discuss!

Before you proceed, discuss the TIPI with a friend or group from class. If you cannot have these conversations in person, use a class discussion board or a group chat on something like Canvas or Blackboard. Talking about your ideas helps you think critically about what you've learned, gives you a chance to hear new perspectives, and allows you to practice communicating your ideas effectively. Communicating effectively is an important skill, whether you are working in a research team, solving problems in a professional setting, or just trying to understand the world better. So, give these questions a shot.

1. Which personality traits would you look for in an ideal coworker, friend, or romantic partner? Do these traits change as you consider different relationships or do they

- remain stable? How do the traits you desire in each person compare to your own personality?
2. How do your scores compare to how you see yourself? Do they align or are there surprises? Which personality trait would you change if you could?
 3. Would your scores on the test be the same if you retook it next week? What about ten years from now? Why might personality change? Why might it stay the same?
 4. Why do you think researchers developed a short measure like the TIPI? What trade-offs might they face compared to using a longer personality test?
 5. Are there any ethical concerns when collecting personality data online? If so, how might researchers address those concerns?

Throughout Part I of this book, the icon above will appear many times. Anytime you see this icon, we want you to add the assignment, discussion questions, or other activities to a Word or Google document that will serve as your research portfolio.. This will be a record of all or your research activities and longer projects. The portfolio will allow you to reflect on what you have learned throughout the semester. In addition to being part of your grade for this class, your portfolio can be used when applying to graduate school or any relevant jobs. By creating a collection of your work, your development as a student and researcher unfolds. Your first chance to create your portfolio begins now.

Portfolio Entry #1 – Personality and the TIPI

Copy and Paste your raw scores and percentile scores from the TIPI onto your portfolio.

Copy and Paste the average scores into your portfolio.

Thought Question: After reflecting on your scores and the discussion questions above, write 2-3 reflection sentences about this project.

From Curiosity to Theory: Building Scientific Knowledge

Module 1.2

Explore how scientists develop theories to explain patterns in data

So far, you have taken a brief personality assessment, entered your data into a spreadsheet, analyzed the data by calculating average personality scores, and discussed what the results mean. But you may be wondering: where did the idea of five personality traits come from in the first place? And why these five traits? To answer those questions, let's examine how researchers developed modern theories of personality. In doing so, you will learn what scientific theories are and the role of theory in the scientific process.

The Making of a Scientific Theory: The Big Five Theory of Personality

In the mid-1930s, Gordon Allport and Henry Odbert began studying personality with a simple but powerful idea: the words people use to describe personality might reveal its underlying structure. This lexical approach, as it became known, assumed that the most important personality characteristics would be captured in everyday language. To investigate their idea, Allport and Odbert combed through Webster's New International Dictionary, recording every word that could be used to describe differences between people. They found about 18,000. Then, they pruned their list into a taxonomy that described things like enduring traits, physical characteristics, and social or emotional states.

Following Allport and Odbert, several psychologists examined whether the original taxonomy could be used to identify the building blocks of personality (e.g., Cattell et al., 1970; Goldberg, 1981; Tupes & Christal, 1992). To do this, they asked large groups of people to rate both themselves and people they knew on many of the words that Allport and Odbert had identified (e.g., Fiske, 1949). For example, a participant might rate themselves on how "talkative," "organized," or "sympathetic" they are. Then, they rated friends and family members on the same traits. By gathering thousands of ratings, researchers were able to examine which words consistently appeared together.

Examining patterns between words revealed something remarkable: even though people used thousands of words to talk about personality, personality characteristics grouped together into

just a few clusters. Eventually, these clusters came to be known as the **Big Five theory of personality** (Goldberg, 1981)—the idea that personality is best organized into five basic dimensions: openness, conscientiousness, extraversion, agreeableness, and neuroticism (emotional stability). According to the Big Five theory, any personality-related word will correlate with at least one of the five basic traits, and these five traits are universal, which means they are consistent across cultures (John and Srivastava, 1999).

Once the Big Five theory was established, researchers created various questionnaires to measure the traits. These questionnaires include the ten item personality inventory you completed earlier as well as the [Big Five Inventory](#) (BFI for short, contains 44 items) and the NEO Personality Inventory with 240 items (Costa & McCrae, 2008). Each of these measures has been translated into different languages and presents a tradeoff between the time taken to administer the test and the precision in the results.

What Makes a Theory Good?

The idea that personality traits cluster into five dimensions is an example of a **scientific theory**. In everyday language, the word "theory" often means a guess or speculation, as in "I have a theory about why my roommate is always late." In science, however, a theory is a data-driven explanation for a set of observations. And theories make specific predictions about what should happen in different situations.

Good scientific theories share several important characteristics (see Figure 1.5). First, they are **data-driven**, which means they emerge from many measurements or observations. The Big Five theory is a good example; its five traits were derived from hundreds of thousands of personality ratings collected by researchers spanning from Allport and Odbert to today.

What Makes a Good Scientific Theory?

Data-Driven	Parsimony	Falsifiability
<p>Definition: Emerges from systematic observation and analysis of large amounts of data</p>	<p>Definition: Explains complex phenomena using the simplest possible explanation</p>	<p>Definition: Makes specific predictions that could potentially be proven wrong</p>
<p>Example: Big Five theory derived from analysis of thousands of personality trait ratings from many studies</p>	<p>Example: Thousands of personality descriptors reduced to just five fundamental dimensions</p>	<p>Example: Prediction that the same five dimensions will be found consistently across different cultures and languages</p>

Figure 1.5. Elements of a good theory.

Second, scientific theories strive for **parsimony**. This means they take something seemingly complex and explain it in the simplest way possible. The Big Five achieves parsimony by showing how thousands of different words can be organized into just five traits. Over the years, scientists have tested whether any of the traits overlap enough to be condensed. If so, three or four traits may capture all the variation in personality, creating a more parsimonious theory. But the data have consistently shown that five dimensions—and no fewer—best account for how personality clusters together.

Third, and perhaps most importantly, scientific theories must be **falsifiable**. This means they must be capable of being proven wrong. The Big Five theory, for instance, makes specific predictions about what patterns researchers should find when studying personality. If these predictions turned out to be wrong, the theory would need revised or replaced.

Theory vs Hypothesis

A theory like the Big Five makes specific predictions—in this case, that five basic dimensions of personality should consistently emerge when studying people from different groups. Researchers translate these kinds of theoretical predictions into specific, testable statements called hypotheses.

A **hypothesis** is a precise claim about what researchers expect to find in a particular study. For example, the Big Five theory predicts that personality has five universal dimensions. Based on this theory, researchers might form the hypothesis that "personality ratings from German participants will reveal the same five dimensions found in U.S. participants." This hypothesis is testable because researchers can collect **empirical data** that either supports or refutes it.

Theories are always subject to refinement. For that reason, scientists avoid saying they have "proven" a theory. Instead, they talk about **evidence** that supports or fails to support theoretical predictions. A pattern of supportive evidence across many studies increases scientists' confidence in a theory, allowing other researchers to build upon its ideas. Evidence that fails to support a theory suggests the ideas need modified.

Testing Hypotheses, Refining Theories

Let's look at how theories operate in the scientific process using the Big Five as an example (Figure 1.6). If personality really has five universal dimensions, then researchers would expect to find the same five dimensions among people from different groups, regardless of language, age, culture, religion, or other factors.

Based on this theory, researchers might decide to see if people from Germany, Brazil, Japan, and Nigeria show a similar five-factor structure in personality, despite their differences in language and culture. To test the hypothesis that personality dimensions should be the same across cultures, researchers would gather data from people in each country.

To do that, they might assemble personality descriptors across different languages, gather ratings from local participants, and analyze whether the same five dimensions emerge. Each time the hypothesis is tested in a new country, the data would either support the Big Five theory by finding the predicted five-factor structure or challenge the theory by finding something else.

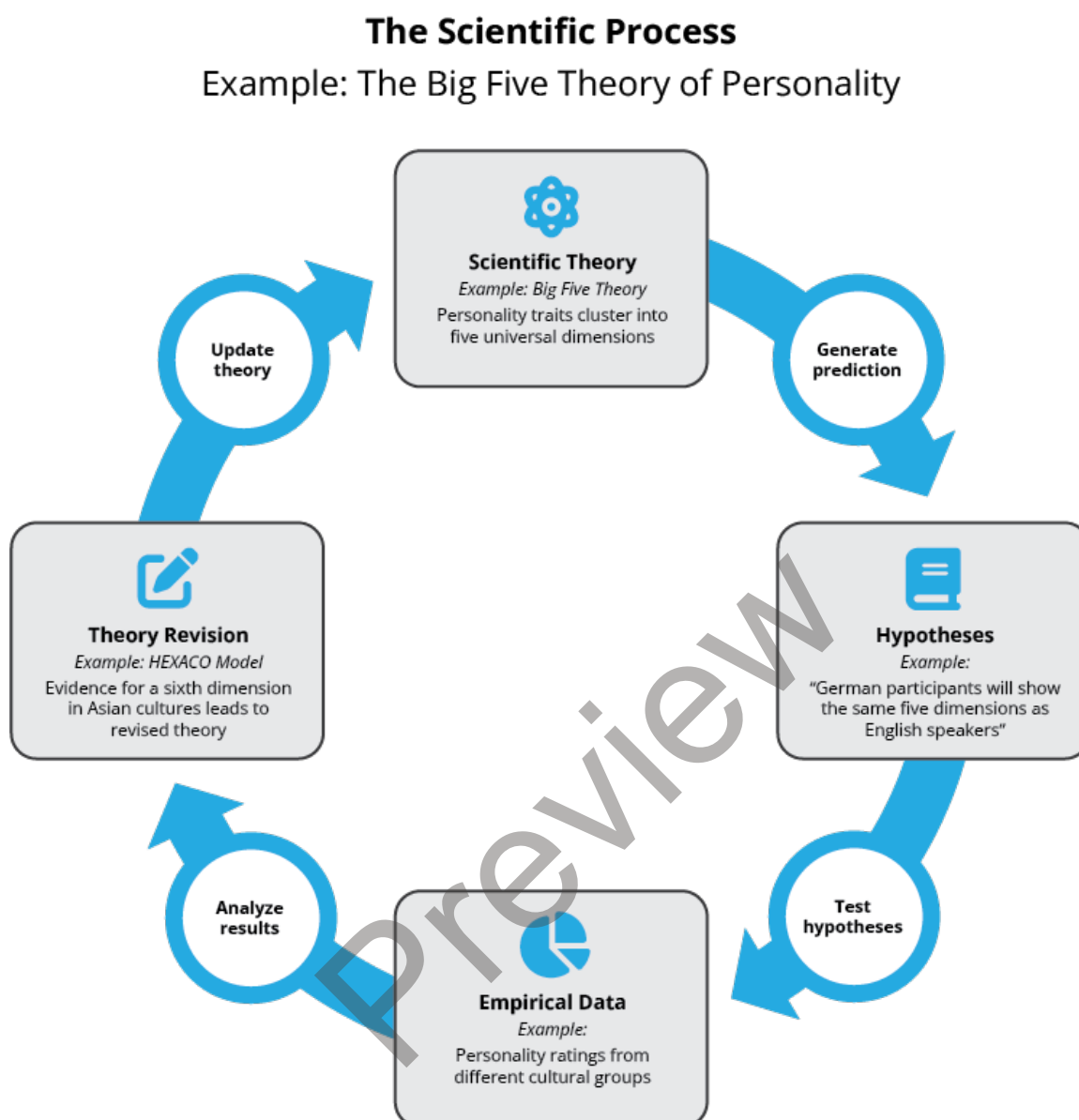


Figure 1.6. A depiction of how theory forms through the scientific process.

Researchers who have done this work have made some fascinating discoveries. Across more than 50 cultures and dozens of countries, the five-factor structure appears, regardless of whether people rate themselves or their peers (e.g., Allik & McCrae, 2004; Benet-Martínez & John, 1998; McCrae & Costa, 1997; McCrae & Terracciano, 2005). However, some studies in Asian cultures have found evidence for a sixth personality dimension related to interpersonal harmony (e.g., Ashton & Lee, 2007; Lee & Ashton, 2004). This sixth dimension has led to an alternative theory called the HEXACO model, which proposes that personality has six basic dimensions instead of

five (HEX means six in Greek). When competing theories emerge, like the HEXACO model, researchers design new studies to test where the theories make different predictions, repeating the scientific process depicted in Figure 1.6.

And this brings us to another aspect of what makes for a “good” scientific theory. A theory doesn't have to be "true" to be useful or influential. Instead, theories are evaluated on how well they make predictions and contribute to scientific progress. Even if a theory is eventually falsified, it can still push scientific knowledge forward by forcing researchers to develop better explanations for the patterns they observe.

Overall, the story of personality theory illustrates how science advances through a continuous cycle of prediction, testing, and refinement. From a general theory like the Big Five, researchers formed specific, testable hypotheses. Then, they collected **empirical data** to test these hypotheses. Each study either supported the theory by finding the predicted result or challenged it by finding something different. When researchers discovered a sixth personality dimension in Asian cultures, it led to a competing theory. Then the cycle continued as researchers gathered more evidence to evaluate which theory better explained the patterns in the data.

Sometimes the evidence clearly supports one theory over another. Other times, competing theories both capture important aspects of reality. For instance, the debate between five versus six personality dimensions continues today (Thielmann et al., 2022), with different studies supporting different conclusions. It is, however, only through this ongoing process of theory development, empirical testing, and refinement that scientific knowledge grows more sophisticated over time.

Behavioral Science in the Real-World

Module 1.3

Examine where behavioral science is practiced

Earlier, you experienced how behavioral scientists measure variables and analyze data. While the project you completed was relatively simple, you did the same thing professional researchers do when they address real-world problems. For instance, organizations often use personality assessments to improve hiring decisions and build effective teams; clinical psychologists use them to understand clients and tailor approaches to treatment; the military uses personality testing to match recruits to roles where they are likely to succeed; and courts use personality assessments when making decisions about rehabilitation programs or evaluating people's mental competency. Even dating apps use personality tests to help people find compatible partners.

Yet, there is much more to behavioral science than the study of personality.

What is Behavioral Science?

Behavioral science is generally understood as the study of how people and animals behave, think, and interact with each other and their environments. It explores the causes of human thought, emotion, and behavior by examining the influence of biology, psychology, culture, and society. Behavioral scientists learn about people using methods like observations, experiments, and surveys and they often apply their knowledge toward solving problems, improving well-being, or designing better systems and policies.

A behavioral scientist might, for instance, be a therapist helping people navigate mental health challenges, a professor researching and teaching consumer behavior, a neuroscientist analyzing brain activity in mice, or a social psychologist studying how group dynamics influence individual decision-making and conformity. They might also be a school psychologist helping students overcome learning difficulties, an economist studying home prices or inflation, a demographer analyzing population trends, or a researcher examining how people use technology. The world of behavioral science is big and it is diverse.

The Disciplines of Behavioral Science

Table 1.1 lists several academic disciplines that fall under the umbrella of “behavioral science.” Examining the table can give you a sense of the field’s breadth. But this list is incomplete. Not every field that examines human behavior has been included, and within disciplines like psychology, there are often sub-disciplines such as cognitive psychology, social psychology, industrial and organizational psychology, and clinical psychology. Further complicating classification, the boundaries between disciplines are often porous. Scientists trained in one discipline may pursue a topic traditionally associated with another discipline. Or, as is increasingly common, researchers with different areas of expertise may collaborate in what is called interdisciplinary research.

Discipline	Description
Anthropology	Explores human cultures, societies, and their development over time.
Behavioral Neuroscience	Studies the relationship between the brain, nervous system, and behavior.
Communication Studies	Investigates human communication patterns and their effects on individuals and society.
Criminology	Examines the causes, consequences, and prevention of crime and deviant behavior.
Economics	Investigates how individuals, groups, and societies allocate resources and make decisions.
Linguistics	Studies language structure, usage, and its role in human communication and cognition.
Marketing/Consumer Behavior	Examines how individuals and groups make decisions about purchasing and using goods and services.
Political Science	Studies political processes, institutions, and behavior within social systems.
Psychology	Examines human thoughts, feelings, and behaviors through scientific methods.

Public Health	Focuses on protecting and improving population health by studying behaviors and social determinants.
Sociology	Studies social structures, relationships, and the patterns of human behavior within groups.

Table 1.1. The Disciplines of Behavioral Science

Note: The behavioral sciences encompass a range of disciplines that explore human behavior, each with unique questions and methods.

Where Behavioral Science Happens: Labs, Businesses, Governments, and Beyond

People who are trained as behavioral scientists work in a variety of organizations. While many work at colleges and universities, combining research with teaching, others work in industry helping organizations understand consumer behavior or improve employee performance. A significant number work at think-tanks and government agencies, evaluating policy or tracking vital population statistics. Others work in healthcare or technology, in non-profits and international organizations, applying their expertise to improve public health, design user-friendly systems, or address pressing social issues.

Given the diversity of behavioral science, you might wonder: what ties these fields together? The answer is a commitment to scientific methods. In fact, the emphasis on careful observation, systematic measurement, and the rigorous testing of ideas isn't bound by location or organization, which means behavioral research can pop up in unexpected places like professional baseball.

Perhaps you are familiar with Michael Lewis's 2003 book *Moneyball*—it was made into a movie starring Brad Pitt. *Moneyball* tells the story of the Oakland Athletics and how they revolutionized player evaluation using scientific principles. Rather than rely on traditional scouting wisdom or a gut feeling about players, the A's scientifically studied what helps teams win. When they applied this knowledge during the 2002 season, they won 103 games—including an American League record at the time, 21 in a row—and tied for the best record in baseball while spending about \$100 million less than their competitors. Eventually, their approach was adopted by other teams, and then it spread to other sports.

What makes science such a powerful approach to human behavior is its ability to reveal unexpected insights and challenge common assumptions. Just as *Moneyball* challenged traditional wisdom in baseball, behavioral science frequently uncovers surprising patterns in human behavior that contradict what “everyone knows” to be true. In the next chapter, we will show you how to use some of the tools researchers use to make these fascinating discoveries.

Summary

In this chapter, you took the first steps into behavioral research by completing a personality test and learning how researchers measure human characteristics. You saw how researchers turn complex psychological traits into measurable variables, how they enter data into databases, and how their findings are applied in various real-world settings. Through the development of personality theory, you learned how scientific theories emerge from careful observation, make testable predictions, and evolve as new evidence emerges. While personality is just one small piece of behavioral science, it illustrates the field’s broader mission: to systematically study human behavior using scientific methods.

You also learned about the essential characteristics of scientific theories. Good theories are data-driven, parsimonious, and falsifiable. Scientists evaluate theories based on how well they make useful predictions and contribute to the development of scientific knowledge. Through a continuous cycle of prediction, testing, and refinement, scientists develop an increasingly sophisticated view of human behavior.

In the next chapter, you will learn about the platforms researchers use to recruit participants, the software they use to design studies, and the methods they employ to analyze data. Most importantly, you will start using these tools yourself. In the next chapter, you will move one step closer toward conducting your own behavioral research.

Chapter 2: Behavioral Science Tools

Every profession has its tools. Carpenters swing hammers, farmers drive tractors, surgeons wield scalpels, and brewers tend fermenters. What tools do behavioral scientists use?

In this chapter, you will learn about the tools that make modern behavioral research possible. With these tools, researchers today collect, analyze, and interpret more data about the human experience than at any other point in history. And, as with Chapter 1, you're not going to just read about these tools—you're going to use them.

In Module 2.1, we will introduce you to online research by having you participate in several studies. Participating in these studies will allow you to see how behavioral research operates and what it looks like from the participant's perspective. Afterward, we will explore the tools that make research possible (Figure 2.1). You will learn how researchers develop ideas and review the existing literature in Module 2.2. You will learn about designing studies and gathering data in Module 2.3, before you learn how researchers analyze data in Module 2.4. Finally, the chapter will end by describing tools researchers use to communicate about their results and share their data. By walking through each step of the research process, you will see how the results of one project often serve as the starting point for another, continuing the research cycle.

Throughout the chapter, we will use an example project to make each step concrete. The project examines personality traits that are associated with navigating life as a single person versus in a committed relationship. The project builds upon the personality research you learned about in Chapter 1 while giving you the opportunity to see how researchers might investigate such a question. So, when you are ready, let's dive into the tools of behavioral research!

The Behavioral Scientist's Toolkit

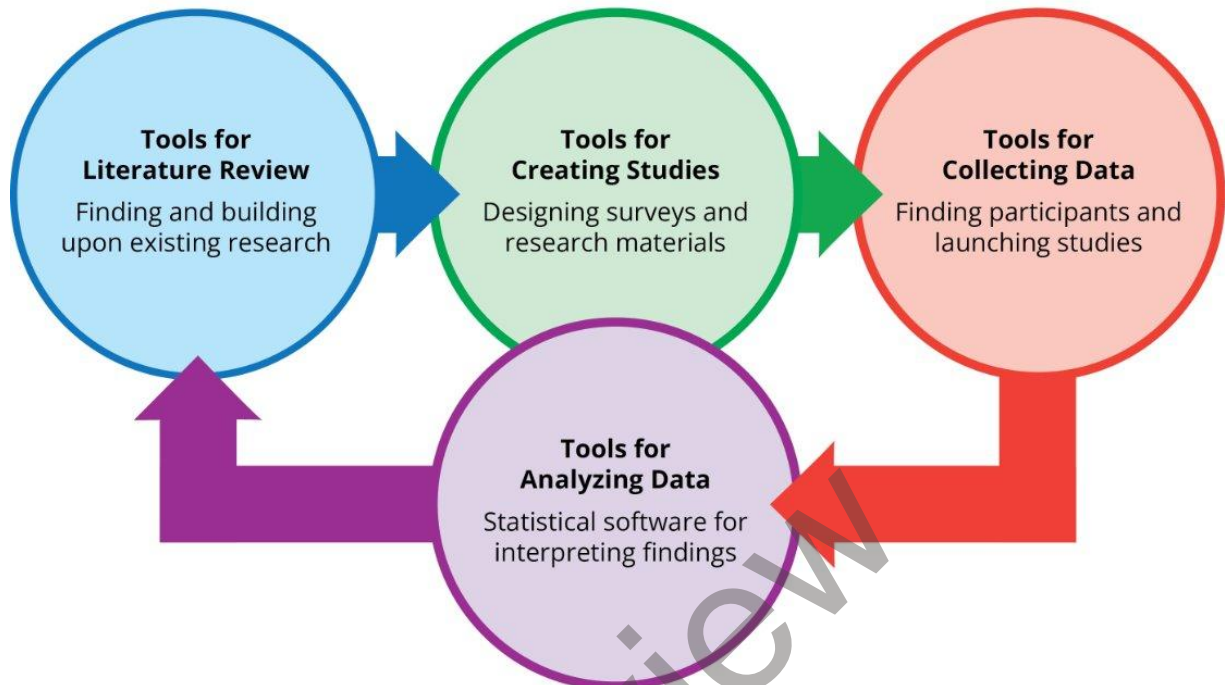


Figure 2.1. Behavioral scientists use different tools at each stage of the research process.

Chapter 2 Outline: Behavioral Science Tools

Module 2.1: Tools for Finding Research Participants



Laboratory

Participate in behavioral science studies and discover the platforms researchers use to recruit participants

- Sign up as a participant in an online research platform
- Learn how researchers find participants by participating in some studies yourself
- Examine the researcher's side of recruitment and the steps of the research process

Module 2.2: Tools for Literature Review



Laboratory

Learn how to find published research

- Explore literature reviews and why they are valuable
- Use Google Scholar to Conduct your Own Literature Review
- Explore the pyramid of knowledge in peer-reviewed research

Module 2.3: Tools for Creating Studies



Reading

Explore the tools researchers use to design studies and measure behavior

- Learn about the tools researchers use to gather self-report, performance, physiological, and behavioral data

Module 2.4: Tools for Analyzing Data



Reading

Examine the tools researchers use to analyze data

- Explore traditional tools of data analysis
- Learn how AI is changing data analysis

Module 2.5: Tools for Sharing Research



Reading

Explore how scientists communicate findings by using platforms that support open science and public engagement.

- Learn how scientists communicate about research
- Learn about the Open Science Framework for sharing data and materials

Tools for Finding Participants

Module 2.1

Participate in behavioral science studies and discover the platforms researchers use to recruit participants

Introduction to Online Research

Behavioral research is about curiosity. Once researchers find a topic they are interested in, they observe people's behavior, form hypotheses, test their ideas in studies, and use data to draw conclusions. This systematic process of discovery occurs in dozens of disciplines and across all kinds of organizations. It also has big consequences. Research helps doctors understand how to treat mental illness; it allows businesses to market products; it helps lawmakers understand crime; and it enables scholars to explain everything from decision-making to discrimination. In other words, curiosity can have a big impact.

Today, most behavioral research doesn't happen in labs or classrooms—it happens **online**. Researchers around the world rely on specialized tools to find participants, conduct studies, and collect data. While you will learn to use these tools throughout this book, for now, we want to introduce online research by focusing on the role of research participants in behavioral science.

The Role of Participants in Science

Behind every discovery in behavioral science there are people who have chosen to provide data by participating in a research study. Historically, these people have been hard to find.

Before the internet, researchers often found participants from local universities or community advertisements. These methods were slow, and they limited who could participate to people physically near the researchers. Then, the advent of online tools transformed how behavioral scientists find participants. Today, there are several digital platforms that connect researchers with participants from all around the world. These websites allow researchers to post studies targeted at specific people. Participants browse projects by topic, time commitment, and compensation, completing the ones that look appealing.

Online research has become so common that more than 80% of published studies in social psychology have at least some participants from online sources (Zhou & Fischbach,

2016). When recruiting participants online, researchers can quickly reach people from different geographic locations, age groups, educational backgrounds, and life experiences. They can also get these people to do things that would be too burdensome offline, such as filling out a survey multiple days in a row.

Although online tools make research easier, it's important to remember that participants are more than data points. They are people who offer their time, attention, and thoughts or behaviors so researchers can test ideas and build knowledge. Without participants there is no study, no analysis, and no scientific discovery.

In a few minutes, you will see how behavioral research operates online by participating in a few studies. You will create a participant account on a commonly used platform operated by CloudResearch, called Connect. Then, you will participate in at least three studies. These studies are posted by researchers from all around the world, spanning all the disciplines and organizations you learned about in Chapter 1. After you complete a few studies, you will step into the shoes of a researcher and learn about the tools that made your experiences possible.

Research Activity 2.1: Participating in Online Studies

You might be thinking: What will I learn from participating in studies? The answer is several things, including some you can't learn any other way.

First, you will learn about the incredible variety of questions behavioral scientists investigate. While you read about this diversity in Chapter 1, participating in studies will allow you to see it. In one study, you might help researchers understand how people make financial decisions, how people form first impressions, or how they solve complex problems. In another, you might participate in a psychology experiment, take a survey about your political views, or play an economic game. You might even be asked to serve as a juror in a mock trial. Each study will give you a look at behavioral research, and unlike the demonstrations in Chapter 1, your data in these studies will help researchers better understand behavior.

Second, you will experience the tools and methods researchers use to collect data online. You might be asked to complete studies with interactive tasks, survey questions, or games and

simulations. Many of these tasks will be presented in survey platforms that you will learn to use later in the book.

Finally, and perhaps most importantly, serving as a participant will make you a better researcher. You will learn what it feels like when instructions are clear (or confusing), when a study is engaging (or boring), and when the compensation seems fair (or not). These insights into what it's like to be a participant will be invaluable when you start designing studies and recruiting participants later in the book. In fact, one of the first steps many professional researchers take when getting started with online research is to participate in studies so they can understand the issues mentioned above.

Plus, did we mention you'll get paid? Most projects on Connect pay between \$10 and \$15 per hour, although the typical project only lasts about 10 minutes. Even so, we are willing to bet you have never had another class assignment that allowed you to earn money while learning!

Creating a Participant Account

To help set up your participant account, we have created a video: [video link]. You can watch the video or follow the instructions below. Signing up takes a few minutes and requires completing an onboarding process, but once it is done you will be ready to participate in studies.

To create a Connect account, navigate to the CloudResearch website: <https://bit.ly/3RoVBnZ>.

Enter your name and college or university email address (e.g., a .edu address within the U.S.). Select the "Research in the Cloud" book for how you heard about the site. Then select "Create Account."

You will be asked to verify your email address. Click on a link sent to your email account to do this. After verifying, you can log back into the site.

When you log in, you will see a screen like Figure 2.2. This screen welcomes you and starts the onboarding process. Onboarding takes a few minutes, but it is an essential part of maintaining high data quality on Connect, which you will learn about in Chapter 10.

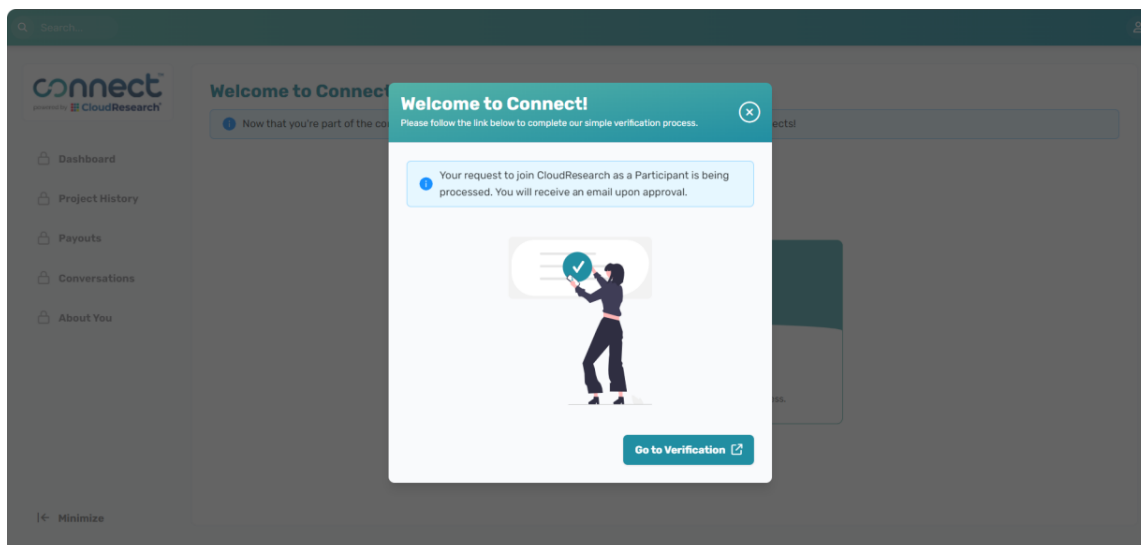


Figure 2.2. When joining Connect, you'll see a welcome message and a link to begin onboarding.

After onboarding, you will be granted access to the site. Before taking studies, you will be asked to provide some basic demographic information and configure a payout method (Figure 2.3). While most participants connect to a PayPal account, you can choose Amazon gift cards or a transfer directly to your bank account.

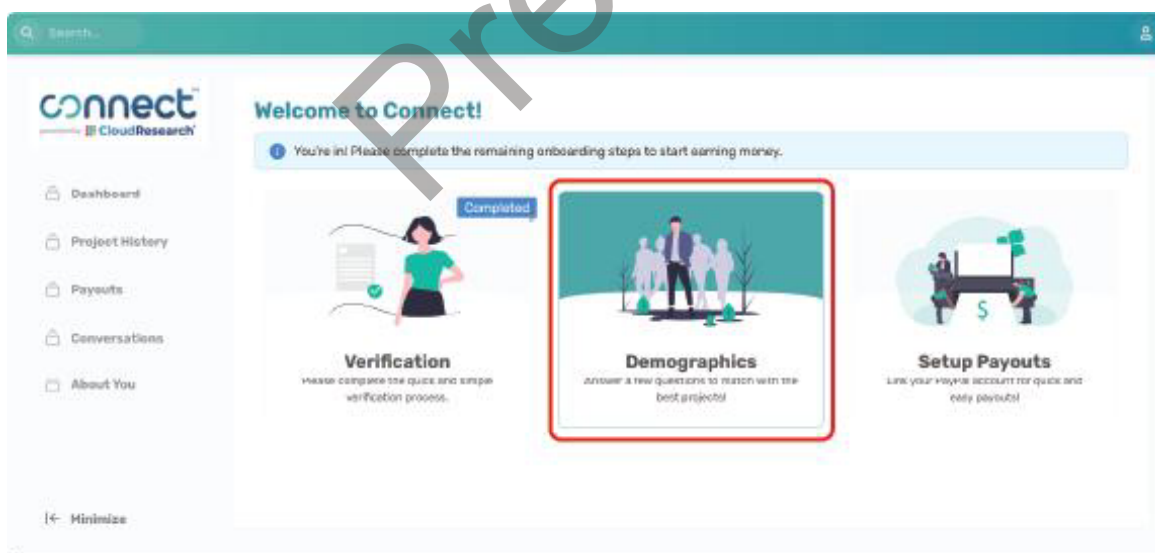


Figure 2.3. After onboarding, participants are asked to answer some demographic questions and then configure a payout method.

Participating in Projects.

The participant dashboard looks like Figure 2.4. From the dashboard you can see available projects.

The projects on Connect are designed by researchers from all over the world. Many of these researchers work at colleges and universities, but others work at multinational corporations, government agencies, non-profits, or private businesses. Some are even student researchers just like you'll be soon!

Using the menu on the left, you can review your project history, cash out your earnings, communicate with researchers, or complete your user profile in the “About You” section.

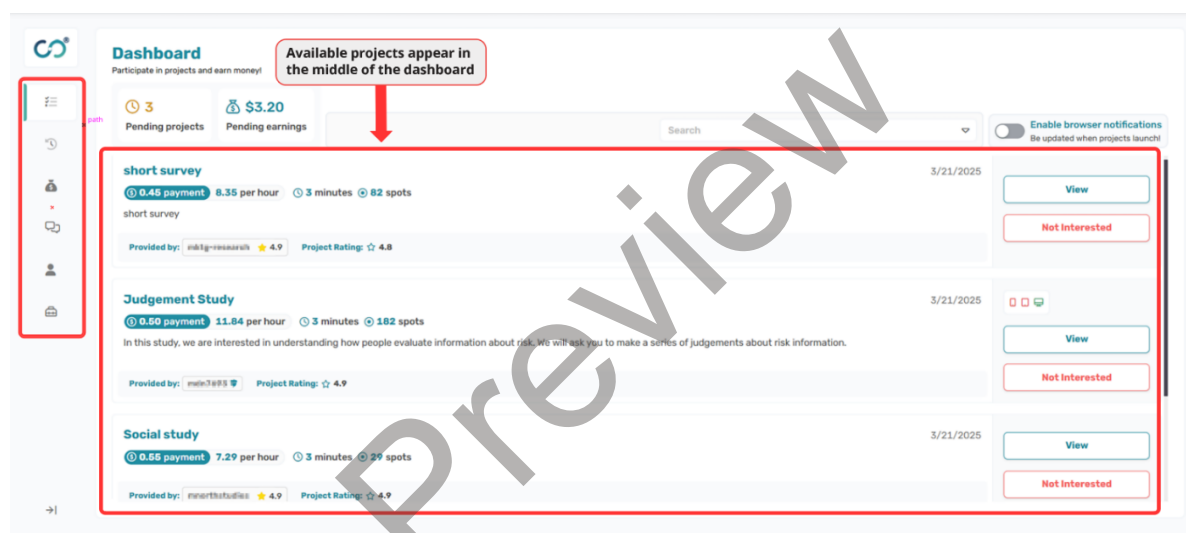


Figure 2.4. The participant dashboard displays available studies.

When you see an interesting project, select “View.” This will take you to a page that summarizes what you will be asked to do and presents any special instructions for the project (Figure 2.5). Special instructions are where researchers communicate technical requirements (e.g., downloading software), warn about sensitive content, or inform participants that the project involves follow-up surveys.

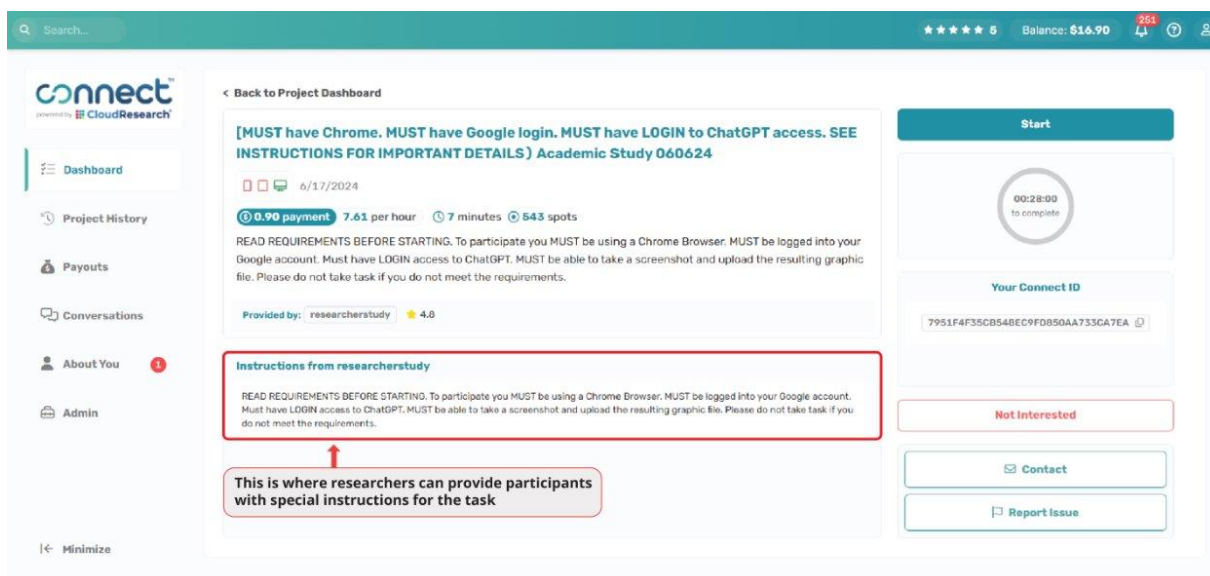


Figure 2.5. The project preview provides participants with information before they accept a study.

To accept a project, click “Start.” You will be presented with a link to the survey. Clicking “Go to Study” will take you out of Connect and to the study created by the researcher. Usually, this study is in Qualtrics or a similar survey platform. Once you complete the project, you will be automatically directed back to Connect or given a completion code to copy and paste into the submission box on Connect.

And that’s all there is to it! You are ready to participate in behavioral research. As a reminder, you should complete at least three projects. In these projects you can find everything from consumer surveys and experimental studies to mock jury trials, interviews with artificial intelligence, and fun personality tests. After you have taken a few studies, come back here to learn about the research process and how these studies got on Connect in the first place.



Research Portfolio

Throughout Part I of this book, the icon above will appear many times. Anytime you see this icon, we want you to add the assignment, discussion questions, or other activities to a Word or Google document that will serve as your research portfolio. If your teacher wants, you can turn

this in as evidence of what you are learning. Even if your class doesn't do that, however, you can keep the portfolio for your own records.

You will have many opportunities to conduct research throughout this book. With some planning and forethought, your projects can build upon each other to investigate a topic or question you are interested in. Then, if you apply to graduate school or a research job down the road, you will have a list of accomplishments you can describe or show to employers. Your first chance to create your portfolio begins now.

Report on Your Experience

After you complete at least three studies, record the information below and be prepared to discuss it with your classmates. For each question, writing a few sentences should suffice. The goal is to help you reflect on your experience and help you start thinking like a researcher.

1. For each study, list the title.
2. How long did each study take?
3. Describe which study you enjoyed the most (or least) and why. Please be sure to provide details of the study and what kind of tasks you were asked to complete. Was this a survey? What kind of questions were you asked? Did you have to provide your opinion or preferences?
4. Generate a research hypothesis (or hypotheses) for your favorite or least favorite study. In other words, what do you think the researchers were trying to learn? Why do you think they designed the study the way they did?

For example, if you completed a study asking about feelings of depression, what kind of instrument did the researchers use to measure it? Which aspects of depression do you think the researchers were interested in learning about? Were you asked any demographic questions? Was the task, easy, hard, fun, or boring?

5. How has participating in research added to your understanding of behavioral science?
6. Is there anything you might do differently as a researcher in any of the studies you completed?

From Participant to Researcher

Now that you have taken part in research, it's time to switch perspectives. We want you to step into the shoes of a behavioral researcher.

Behind each study you completed was a researcher, or team of researchers, who asked the question, designed the study, chose the tools, collected the data, and figured out what the results meant. As you transition from a participant to a researcher you should shift from thinking, “*What’s this study about?*” to asking “*How can I create this study myself?*” Being a researcher means thinking critically about how to measure behavior, how to find good information, and how to tell a story with your results.

To get started, you can create a Connect researcher account. This will allow you to create studies and collect your own data down the road.

To create an account, go to: <https://account.cloudresearch.com>. Make sure you choose "Connect for Researchers" (Figure 2.6) and use the same email and password you used as a participant. Once you have logged in, you are officially a researcher on Connect. With this account, you will be able to create and launch studies like those you just participated in.

The screenshot shows the CloudResearch Connect login interface. On the left, under '1. Select your product:', the 'Connect' option for researchers is highlighted with an orange border and a red arrow. On the right, under '2. Sign in:', there are input fields for email and password, a 'Remember me' checkbox, and a 'Forgot Password?' link. A blue 'Log in' button is positioned below these fields. To the right of the sign-in fields are social media login buttons for Google, Facebook, Microsoft, Amazon, and Twitter. Below the sign-in fields is a link for 'New to CloudResearch? Create an account.' and a disclaimer at the bottom.

Figure 2.6. You can log into the researcher side of Connect with the same credentials as the participant side.

The researcher side of Connect looks like Figure 2.7. From this screen, you can use the icons on the left to communicate with participants, create studies, and manage participants.

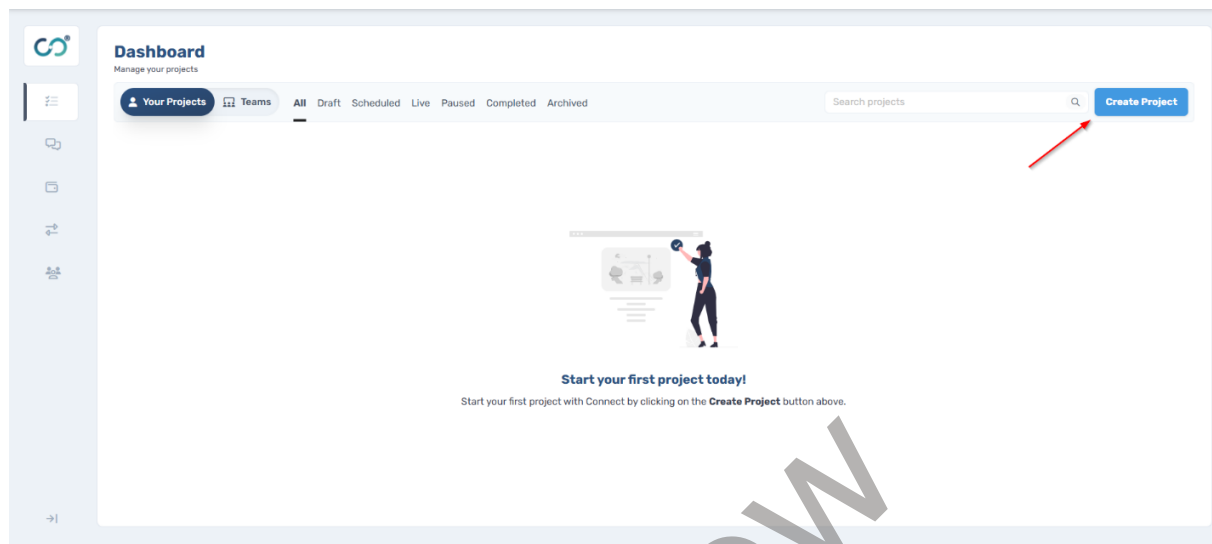


Figure 2.7. Click "Create Project" to set-up a new study.

If you click the "Create Project" button, you will enter the study set-up window. From there, you can prepare a project by describing what participants need to do, pasting a link to your survey, determining who is eligible based on hundreds of demographic characteristics, and making decisions about how many people you want to recruit. Once you launch the project, it is available to participants.

Throughout the rest of this book, we will show you how to set up online projects, recruit participants, and gather data that answers questions you are curious about. But first, we need to walk through the research process. This is where research gets exciting. So, let's begin.

The Research Cycle

As you participated in studies, perhaps you wondered: Who is conducting this research? How did these projects come to be? What happened before this study appeared online? And what will the researcher do with the data afterward?

These questions are about the research process, which begins long before participants see a study and continues after they have completed it. Let's outline this process before exploring the tools that make it possible.

Step 1: Generating ideas and reviewing the literature. Research begins with curiosity and a research question. But before designing a study, researchers need to know what others have already discovered about the topic. Specialized search tools like Google Scholar make it possible to quickly search the published literature in a scientific area. Reading this literature helps researchers develop questions and connect their ideas to existing theory.

Step 2: Designing studies and collecting data. Next, researchers must choose how to measure the things they are interested in. Will they use surveys? Cognitive tasks? Physiological measurements? Will the study take place online or in person? For many research questions, online survey platforms like Qualtrics provide a powerful and user-friendly way to design professional surveys.

Step 3: Finding participants. Once the study is designed, researchers must find the participants. Online platforms like Connect make it possible to find diverse participants in hours rather than weeks or months.

Step 4: Analyzing data and visualizing the results. After collecting data, researchers use statistical tools to analyze patterns and create visualizations that help communicate the findings. These tools transform raw numbers into conclusions.

Step 5: Sharing findings. Finally, researchers share what they have learned. Once the findings of a study are disseminated to others through publications and presentations, they become part of the scientific record. As other scientists review the work and form new questions, the research cycle begins again.

In the rest of this chapter, we focus on the tools researchers use to develop ideas, design studies, analyze data, and share their results. As you just learned, the process begins with a literature review.

Tools for Literature Review

Module 2.2

Learn how to find published research

Where do research ideas come from? Curiosity is the simplest answer. Whenever you observe people around you, notice trends in society, or learn about behavior from books, movies, and media, you might think: why do people do *that*? Anytime you have this experience, you have an idea for research.

Let's imagine you are interested in the Big Five personality traits described in Chapter 1. Let's also imagine you notice surveys and other data showing an increasing number of people are moving through life without committing to a long-term partner. In other words, they are single. You decide to test whether people who are single have different personality characteristics than those in relationships, and whether one group of people is happier than the other.

Before you read further, look at the Big Five traits in Chapter 1 (see Box 1.1). Create some hypotheses about what the personality differences might be between people who stay single versus those who commit to long-term relationships. Write these down in your portfolio.

Which personality traits do you think are more associated with staying single versus having a partner? Who do you think is happier and why? Discuss your ideas with your friends or classmates. Later in this chapter, you will see if your hypotheses are correct by comparing them to what research has found.

The Value of a Literature Review

When researchers are curious about a question, one of the first things they do is explore what is already known about the topic.

Has anyone investigated this question? If so, what did they find? Perhaps previous research has only looked at the question among college students but not older adults (or the reverse).

Understanding what research has already been done allows scientists to design studies that build upon existing knowledge.

To find existing studies on a topic, researchers conduct a **literature review**. Reviewing research in an area helps scientists avoid duplicating others' work. It also allows them to find holes in existing knowledge and connect their research to broader theories in the field. Perhaps most importantly, reviewing previous research helps generate hypotheses that are based on patterns and findings already established in the field.

Imagine what it would feel like to design a study on personality and relationship status, gather the data, and then discover that dozens of similar studies already exist. Not only would you waste time and resources, but you would miss the opportunity to learn something new.

Research Activity 2.2: Using Google Scholar to Conduct a Literature Review



Google Scholar (scholar.google.com) is a powerful tool that allows researchers to search for scholarly articles, books, conference papers, and other academic resources. Google Scholar provides a simple, user-friendly interface that works similarly to the regular Google search engine.

For this activity, you will use Google Scholar to explore the existing research on personality differences between single people and those in committed relationships. This activity is meant to show you the basics of a literature review.

To conduct a search on Google Scholar, start with some basic search terms such as "personality trait differences between singles and partnered people." Or, you may use "personality differences between singles and non singles". When you search these terms, Google Scholar will return a list of relevant articles, ranked by factors like citation count (how many other papers have referenced the article), publication date, and relevance to the search terms (Figure 2.8).

Google Scholar search results for "personality trait differences between singles and partnered people". The search bar shows the query. The results list several articles. The second article, "Differences between lifelong singles and ever-partnered individuals in Big Five personality traits and life satisfaction" by J. Stern, M.D. Krämer, and A. Schumacher (2024), is highlighted with a red box. A red arrow points to the "Cited by 32" link for this article. Other articles include "Similar to and/or better than oneself? Singles' ideal partner personality descriptions" and "Differences between lifelong singles and partnered individuals in personality traits and life satisfaction".

Figure 2.8. Search results on Google Scholar.

Near the top of the results, you should see a 2024 study by Julia Stern and her colleagues titled “Differences between lifelong singles and ever-partnered individuals in Big Five personality traits and life satisfaction.” If you read this paper, you will see that the research examined data from thousands of European participants over the age of 50 and found some interesting differences in people’s personality traits. We encourage you to read the abstract to see whether your hypotheses align with their findings.

A useful feature of Google Scholar is the “Cited by” option that appears in the search results (see the red arrow in Figure 2.8). The “Cited by” feature allows you to see papers that have referenced the study you are looking at. Reviewing these papers helps find related research because a study that is cited by other papers usually helped lay the groundwork for research on that topic.

You can also use Google Scholar for more advanced searches. By placing quotation marks around phrases like “Big Five personality” you can search for the exact term or phrase in your parentheses. In addition, you can add terms like “longitudinal” to find studies that track personality and relationship status over time. You can also restrict your search by year (e.g., since 2018) or by authors who do research on a topic.

Beyond Google Scholar

While Google Scholar provides access to a lot of literature, you may encounter paywalled content. When this happens, your school's library resources can be useful. Most academic libraries subscribe to many journals and databases. These databases provide students and faculty with access to articles that would otherwise be too expensive to purchase individually.

Commonly used institutional databases include PsycINFO (for psychology research), PubMed (for biomedical research), and Web of Science (for interdisciplinary research). These databases cover specific bodies of literature, and your university library likely provides access to these databases, along with tutorials on how to use them effectively.

Connecting to Peer-Reviewed Research

As you saw in Chapter 1, scientific theories develop through systematic testing and refinement. In peer review, manuscripts submitted to academic journals are evaluated by other experts in the field (professional peers) who assess the quality and validity of the research before it's published.

When conducting a literature review, researchers typically prioritize peer-reviewed sources because they represent the most rigorously evaluated research. This does not mean peer-reviewed research is perfect—scientific knowledge continues to evolve as new evidence emerges—but it does provide a foundation of methodologically sound studies on which to build.

By starting your research projects with a thorough literature review, you connect your work to the broader scientific conversation about a topic. Your study becomes not just an isolated investigation, but a meaningful contribution to the collective understanding of human behavior.

The Pyramid of Knowledge

When doing a literature search, you are likely to come across multiple types of scientific publications, from research articles that focus on one specific question to whole books that provide an overview of an entire field. These different sources fit within the larger structure of scientific knowledge. This structure can be viewed as a pyramid, with each level building upon information from the levels below it (Figure 2.9).

The Pyramid of Knowledge

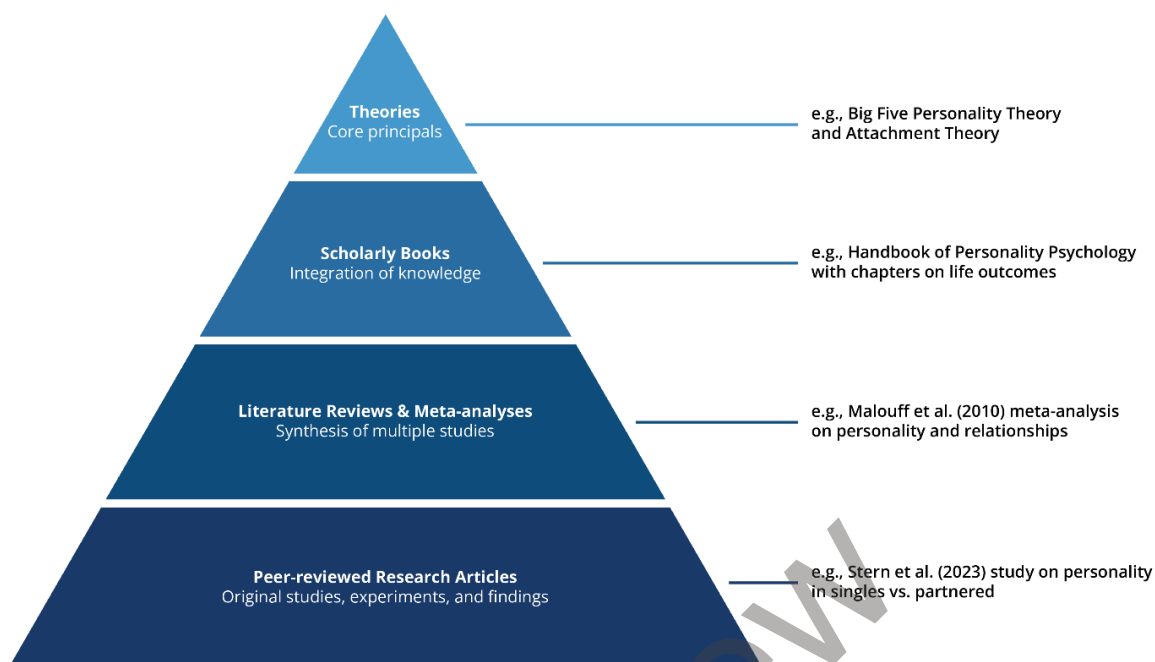


Figure 2.9. Scientific knowledge builds from individual peer-reviewed articles at the base through increasingly synthesized forms of research at higher levels.

At the base of the pyramid are individual peer-reviewed articles. These articles are studies that examined a specific research question. The study by Stern and colleagues (2024) comparing personality traits between lifelong singles and people in relationships is a good example. Each peer-reviewed article adds one block to the foundation of knowledge.

Moving to the second level, you will find articles that synthesize research. These articles are often literature reviews and meta-analyses. These papers systematically gather and analyze findings from many individual studies. A literature review typically summarizes the findings from dozens of studies on a specific topic, whereas a meta-analysis uses statistical techniques to combine the results from dozens or even hundreds of studies. For instance, a meta-analysis by Malouff and colleagues (2014) analyzed data from multiple studies to examine the role that emotional intelligence plays in relationship satisfaction.

The third level of the pyramid contains books, edited volumes, and textbooks. These sources integrate information from review articles, meta-analyses, and individual studies and organize the knowledge into coherent frameworks. A textbook on personality psychology, for example,

might include a chapter on how personality affects major life outcomes including relationships and happiness. While these sources may lack the details found in peer-reviewed journal articles, they provide a more complete understanding of a field.

Finally, at the apex of the pyramid are the major theoretical frameworks and paradigms that shape entire disciplines. In our example, this might include the Big Five theory of personality or broader theoretical perspectives on human relationships. These overarching theoretical frameworks are developed from the knowledge contained across all levels of the pyramid and can be found in every type of publication, including papers, overview articles, and books. Together, the levels of the pyramid contain all the scientific knowledge about a topic.

Preview

Tools for Creating Studies

Module 2.3

Explore the tools researchers use to design studies and measure behavior

After forming an idea and reviewing the literature, researchers design a study. This requires understanding the tools available to gather data. Generally, these tools fall into four categories: tools for gathering self-reported data, tools for measuring cognitive performance, tools for measuring physiological reactions, and tools that track behavior (Figure 2.10).

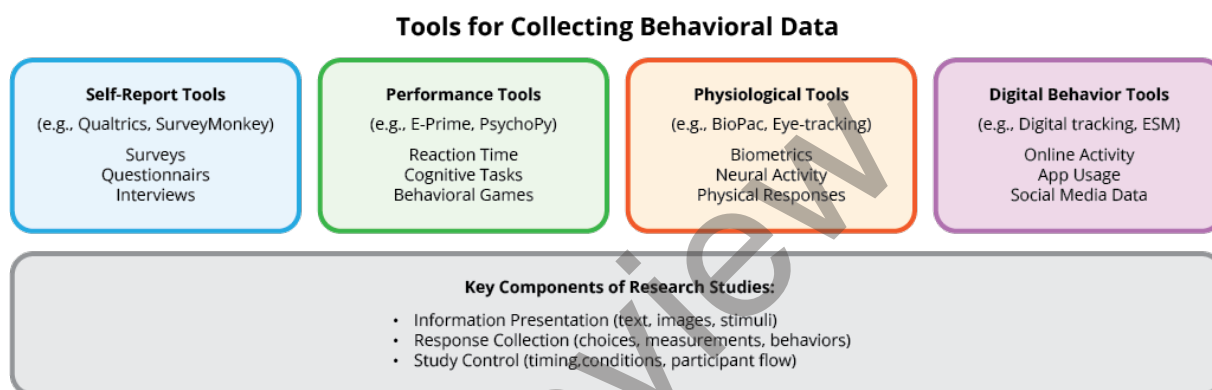


Figure 2.10. Tools for creating four different types of studies.

Survey Platforms: Gathering Self-Reported Data

At the heart of many behavioral studies are survey platforms like Qualtrics, SurveyMonkey, and Google Forms. These digital tools have largely replaced the paper-and-pencil questionnaires of the past and become the workhorses of modern behavioral research.

Survey platforms allow researchers to create professional-looking questionnaires without any programming knowledge. To conduct a study on personality differences and relationship status, we could easily use a platform like Qualtrics to create a study with the Ten-Item Personality Inventory and questions about relationship status. Participants could complete this survey from anywhere with an internet connection, on a computer or mobile device. With the data in hand, we could test our ideas.

What makes these platforms so useful is their flexibility. They can present participants with various types of questions, ranging from multiple choice items ("I see myself as extraverted, enthusiastic" 1 = Strongly Disagree, 7 = Strongly Agree), to open-ended text responses (In your own words, please describe what you think influences your relationship status?). They also allow researchers to present pictures, video, and audio stimuli, and control when these different materials are presented to participants. This flexible stimulus control allows researchers to create sophisticated studies that go beyond what people typically think of as a "survey." In Chapter 14, we provide information about how to create sophisticated research projects using Qualtrics. Before that, virtually every chapter in this book demonstrates some basic techniques of working with Qualtrics.

Figure 2.11 shows what programming the TIPI questions looks like for a researcher in Qualtrics; Figure 2.12 shows how the questions appear to participants. It doesn't take much effort to present participants with professional looking surveys.

The screenshot shows the Qualtrics interface for programming a TIPI question. The left sidebar contains settings for the question type (Likert), answer type (Allow one answer), number of statements (10), and scale points. The main area displays the question text and a table for rating personality traits.

TIPI

Here are a number of personality traits that may or may not apply to you. Please write a number next to each statement to indicate the extent to which you agree or disagree with that statement. You should rate the extent to which the pair of traits applies to you, even if one characteristic applies more strongly than the other.

I see myself as:

	Disagree strongly	Disagree moderately	Disagree a little	Neither agree nor disagree	Agree a little	Agree moderately	Agree strongly
Extraverted, enthusiastic.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Critical, quarrelsome.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Dependable, self-disciplined.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Anxious, easily upset.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 2.11. A view of the matrix question for measuring the Big Five personality traits in Qualtrics.

Here are a number of personality traits that may or may not apply to you. Please write a number next to each statement to indicate the extent to which you agree or disagree with that statement. You should rate the extent to which the pair of traits applies to you, even if one characteristic applies more strongly than the other.

I see myself as...

	Disagree strongly	Disagree moderately	Disagree a little	Neither agree nor disagree	Agree a little	Agree moderately	Agree strongly
Extraverted, enthusiastic	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Critical, quarrelsome	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Dependable, self-disciplined	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Anxious, easily upset	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Open to new experiences, complex	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Reserved, quiet	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sympathetic	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>



Figure 2.12. How the matrix question measuring the Big Five appears to participants.

Beyond basic questions, survey platforms offer advanced features for executing sophisticated studies. The most important of these features is the ability to randomly assign participants to different conditions within a study and to randomize the order of stimuli. This makes it possible to conduct experimental research. For example, participants can be randomly assigned to watch different videos to assess whether they have a different emotional impact, and many other examples we will explore in Chapter 7.

Easy use and flexibility makes survey platforms extremely popular among behavioral scientists. When you encounter a research article reporting findings from a survey or experiment, the chances are the researchers used one of these platforms to gather their data. In addition, some

survey tools are beginning to integrate AI for even greater flexibility and sophistication. Chapter 8 examines one of these platforms and the doors it opens for research.

Performance Tools: Measuring Cognitive Processes

While survey platforms excel at gathering self-reported data about attitudes, beliefs, and experiences, behavioral scientists often need to measure aspects of human behavior that people cannot easily report. To do this, they use specialized software that measures cognitive performance (Figure 2.13). Software packages like E-Prime and SuperLab allow researchers to present stimuli to participants and measure their responses with millisecond precision.



Figure 2.13. The tools that measure cognitive processes are often specialized forms of software that are run on the computer.

For instance, a researcher who is interested in how quickly people can identify emotions might design an experiment where pictures of people with different facial expressions flash on a computer screen for just 200 milliseconds ($1/5^{\text{th}}$ of a second). Participants may be asked to press a key to indicate whether the face showed happiness, anger, fear, or surprise. The researcher wants to know how quickly people process the information and whether they are accurate. Can

you imagine the researcher asking people to self-report how fast they recognized the emotions? Most people would have no clue. For that reason, behavioral scientists don't ask. They measure.

Performance measures are commonly used in fields like cognitive psychology and cognitive neuroscience to study mental processes such as attention, memory, and decision-making. Although these tools require more technical knowledge than survey platforms, they are necessary for many research questions.

For example, we might use these tools to explore whether single and partnered people differ in how quickly they recognize emotional expressions—an aspect of social cognition that could both influence and be influenced by experiences in relationships. Such a study could provide a more complete picture of the psychological differences associated with relationship status.

Physiological Measurement

Beyond what people say or how they perform on cognitive tasks, their bodies provide a wealth of information about what they are thinking, feeling, and doing. Behavioral scientists refer to measures of bodily response as 'physiological measurements,' and these measures provide access to things happening in the body that are often outside of conscious awareness (Figure 2.14).

Physiological research relies upon equipment such as wristbands, heart-rate monitors, eye-trackers, skin conductance devices, and software (e.g., BioPac, Empatica) to measure and interpret people's reactions to different situations. To see the value of these measures, imagine a researcher studying stress in social situations. While participants might report feeling "a little nervous" during a job interview, their bodies may tell a more detailed story. Measuring physiological responses, the researcher might see that the person's heart rate, blood pressure, and skin conductance (how much the skin sweats) are equivalent to levels seen during light physical exercise, such as walking briskly up a flight of stairs. Even though the person's biological responsiveness may show several signs of stress, they can be subjectively interpreted as "a little nervous."



Figure 2.14. An electroencephalograph, or EEG, allows behavioral researchers to measure the electrical activity of neurons within the brain. It is an example of physiological measurement.

In fields like biological psychology and neuroscience, physiological measurement tools are critical for understanding how people's emotional and physiological states interact to influence thought and behavior. In a study on personality and relationships, for instance, we might use physiological measures to examine whether singles and people with a partner differ in physiological reactions when viewing images of couples or thinking about past relationships. Perhaps people with certain personality profiles show a distinct physiological reaction when processing social or emotional information—patterns that might influence their relationship experiences.

While these tools provide valuable data that cannot be collected through other means, they also present some challenges. The equipment can be expensive and requires expertise to use properly. Learning to use these tools typically requires a specialized course in biological psychology research methods and experience working in a research lab. These tools are a key aspect of behavioral research and can often complement the data gathered through surveys and performance measures.

Behavioral Measurement: Capturing Real-World Interactions

The final category of tools track how people behave in their everyday lives. Behavioral scientists are often interested in how people behave in their natural environments. Behavioral measurement tools allow researchers to observe and record real-world behaviors as they occur. One example comes from researchers studying global migration patterns, who used Meta to analyze anonymized data from billions of Facebook users worldwide (Figure 2.15). By tracking changes in users' location data over time, they created detailed maps of migration flows between countries.

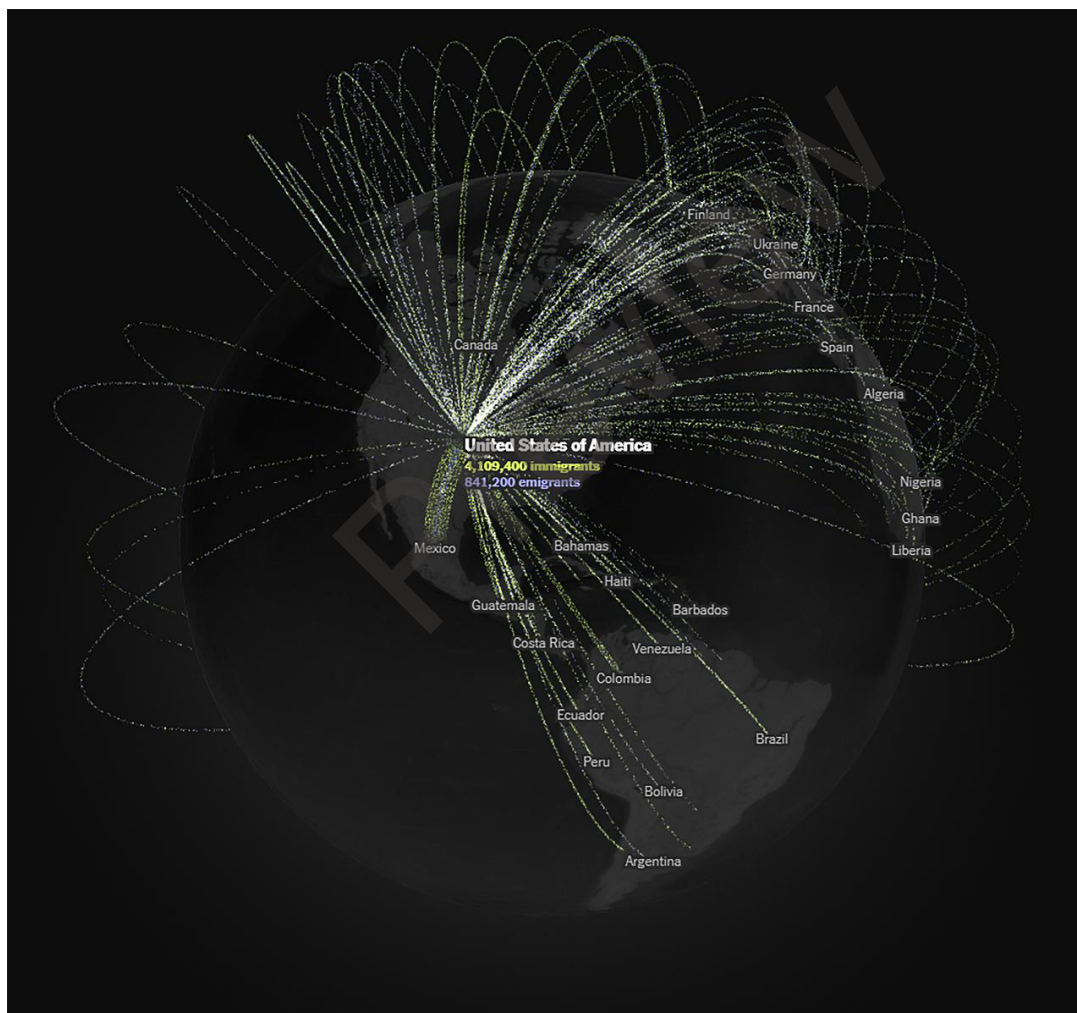


Figure 2.15. Measurement of real-world behavior via Facebook location-tracking reveals global migration patterns (Kingsbury, 2025).

Measuring behavior extends beyond digital tracking. Economists analyze purchasing records to study consumer behavior. Urban planners use traffic cameras to study pedestrian movement. Environmental psychologists place sensors in homes to measure energy usage, as you will see in Chapter 3. In each case, researchers directly measure behavior rather than asking people to report what they did.

Similarly, social media sites, apps, and wearable devices can track people's behaviors across a variety of contexts and provide behavioral scientists with all kinds of data they could never collect in a laboratory. During the early stages of the COVID-19 pandemic, for instance, researchers tracked searches for symptoms like "loss of smell" and "loss of taste" to identify potential outbreaks before official numbers were reported (Cherry et al., 2020). In many cases, behavioral data can be compared to people's self-reported data to see if the two align. If someone says in a survey that they "occasionally check their phone throughout the day" but an app reveals that they actually check 150 times in an average day, the person is either unwilling or unable to accurately report on the behavior.

Making sense of so much data often requires advanced training in statistical analysis. This book will help you develop a foundation you can build on later. When thoughtfully implemented, behavioral measurement tools complement other research approaches by providing objective data about what people actually do.

Tools for Analyzing Data

Module 2.4

Examine the tools researchers use to analyze data

After the data for a study are collected, the next step is *always* to analyze it with statistical tests. In other words, statistical analysis is universal in research.

Traditional Data Analysis Tools

In Chapter 1, you saw the basics of data analysis when you entered your TIPI scores into Google Sheets. The spreadsheet calculated simple statistics like the average score for each personality trait. While spreadsheet programs like Excel and Google Sheets work for basic calculations, most behavioral research requires more advanced statistical tools. These tools rely on specialized software capable of performing complex calculations on large datasets.

One of the most widely used programs in the behavioral sciences is SPSS (Statistical Package for the Social Sciences). SPSS is particularly popular in introductory statistics and research methods courses because of its intuitive user interface (Figure 2.16). In addition to the rows and columns of data you can find in any spreadsheet, SPSS has menus and icons across the top that make it easy to perform all kinds of data operations and analyses. Throughout this book, we use SPSS to analyze data. You will get your first crack at SPSS in Chapter 3.

Another powerful tool is R, a free and open-source program that offers tremendous flexibility for statistical analyses. Simpler alternatives like JASP and Jamovi are often used to teach students the basics of data analysis. They combine an intuitive user interface with powerful statistical capabilities that include built-in explanations of statistical tests, making them excellent for beginners.

Regardless of which statistical package a researcher chooses, the goal remains the same: find meaningful patterns in the data that answer the research question. If we were examining a dataset to look for differences in Big Five traits across relationship status, we might first create an average score for each participant across the five personality traits. Then, we would divide participants into either a singles group or a partnered group (anyone who had ever committed to

a lifelong partner). After preparing the file, we would perform a statistical test to determine whether the average differences between the two groups were meaningful. If we found “statistically significant” results, we would share them with other researchers, as Stern and colleagues did in their paper (2024). They found that in addition to having lower levels of life satisfaction, lifelong singles were less extraverted, less conscientious, and less open to experience compared to partnered people. How do their results align with your hypotheses?

Each column represents a variable from the Big Five

Each row represents a participants' data

Student	Openness	Conscientiousness	Extraversion	Agreeableness	EmotionalStability
Student 1	3.0	4.3	1.3	3.7	2.3
Student 2	5.0	6.5	6.0	4.5	3.0
Student 3	6.5	2.0	5.5	5.5	3.0
Student 4	6.0	5.0	6.0	5.0	5.0
Student 5	4.0	5.0	1.5	4.5	3.5
Student 6	5.5	4.5	3.5	5.5	3.5
Student 7	5.5	5.5	5.0	5.0	2.0
Student 8	7.0	5.5	4.5	7.0	3.0
Student 9	5.0	7.0	3.5	4.5	4.0
Student 10	6.5	2.5	2.5	6.5	5.0
Student 11	5.0	3.0	3.0	5.0	6.0
Student 12	5.0	4.0	4.0	3.0	5.0
Student 13	6.0	7.0	2.0	5.5	3.5
Student 14	5.5	6.5	6.0	4.5	3.5
Student 15	6.5	6.0	6.5	5.0	6.0
Student 16	4.5	6.0	3.0	5.5	4.0

Figure 2.16. The data view in SPSS. Across the top are menus that make analyses easy to conduct.

Statistical packages are not only useful for making sense of data. They also help researchers create visual ways to communicate the results. Charts, graphs, tables, and figures are an important way to effectively communicate the main findings of a research study. In later chapters, you will learn how to use these data analysis tools to test specific hypotheses, create visualizations, and draw conclusions from your data.

AI-Powered Tools for Data Analysis

In recent years, artificial intelligence has transformed many aspects of research, including data analysis. Modern AI tools offer specialized features that can perform sophisticated statistical

analyses. One notable example is ChatGPT's Data Analyst feature, available in the paid version of the platform.

What makes AI-powered analysis tools particularly appealing is their ability to understand commands in everyday language. For example, you might upload a dataset and type, "Calculate the mean scores for each personality trait, create a bar chart comparing them, and conduct an ANOVA to see if the groups are different from each other." Then, like magic, the AI will execute the commands.

If you have access to the paid version of ChatGPT, you can download the Google Sheet from Chapter 1 as an Excel file, upload it to the Data Analyst, and ask it to calculate averages for each personality trait. Then you can ask it to create a bar chart. The result should match what you saw in the original spreadsheet, but you would achieve it through conversational commands.

The Data Analyst feature can handle all the analyses we will cover in this book. We have verified that its output matches SPSS results across all our example projects, from basic descriptive statistics to more complex inferential statistics.

There are several advantages to using AI for data analysis. The conversational interface reduces the learning curve typically associated with statistical software. You can iterate quickly, refining analyses based on initial results and—most importantly—the AI can also explain statistical concepts and answer your questions as it works!

However, as with all AI applications, there are important considerations. AI-powered analysis tools should be used in collaboration with knowledgeable professionals, such as your course instructor. Traditional tools like SPSS provide more transparent workflows that make it easier to verify the accuracy of results, particularly for advanced analyses. The AI might also select analytical approaches that seem reasonable but don't align with best practices for your specific research question.

Throughout this book, we focus primarily on SPSS while acknowledging that AI alternatives exist. We encourage those with access to Data Analyst or similar tools to explore them as complementary resources, especially when getting started with a new analysis or seeking to understand statistical concepts in plain language.

Tools for Sharing Research

Module 2.5

Explore how scientists communicate findings by using platforms that support open science and public engagement.

The work of modern behavioral research doesn't end once the data are analyzed. In fact, it is only then that the work of clearly communicating the findings begins.

Behavioral scientists often communicate about research in public lectures, conference presentations, and journal articles. Common tools like PowerPoint and word processors help researchers organize and disseminate their findings. Chapter 16 provides some advice about writing journal articles.

In addition to publishing research findings, researchers often use platforms like the Open Science Framework (OSF) to organize, store, and share their data and materials (Figure 2.17). The OSF represents an important shift in how behavioral science operates. In the past, researchers typically shared only their results through published articles. The raw data, survey questions, analysis code, and other materials that went into the research remained private. But this practice created problems.

The most important problem was that other scientists often could not easily verify published findings. This became known as the **replication crisis**. The open science approach helped address this problem by providing an online platform where researchers can share their work. Scientists upload their materials, data, and analysis scripts to OSF. These materials become available to anyone interested in understanding or extending the research. This openness promotes scientific transparency and enables other researchers to replicate findings.

Throughout this book, you will access research materials from the *Research in the Cloud* OSF page. You will find survey files you can import directly into Qualtrics, datasets you can analyze, and scripts showing exactly how we conducted our analyses. Using OSF will give you hands-on experience with another one of the online tools that define modern behavioral research. When you work with materials on OSF, you are participating in the culture of open science.



Figure 2.17. The Open Science Framework supports the research process, from start to finish.

Finally, beyond the OSF, behavioral scientists use tools that help them submit research protocols for ethical review; track references and citations of a manuscript; submit grant applications and journal articles for review; and communicate with other researchers while collaborating on a project. None of these tools are essential, but they support the research process and make tasks more efficient.

Summary

In this chapter, you have taken the next steps toward becoming a behavioral researcher. You experienced what online research is like by participating in studies on Connect. Then, you created a researcher account on Connect, taking you closer to launching your own projects. And, finally, you walked through the steps of the research process while learning about the tools researchers use in their work.

Collectively, the integration of so many advanced tools has changed how behavioral research is conducted. A single study might use recruitment platforms like Connect to find participants,

survey tools like Qualtrics to create the study, statistical software like SPSS to analyze the data, and visualization tools to present the findings. Then, the research team might use a data repository like OSF to make their materials and data available to others while issuing a pre-publication version of their manuscript while they await feedback from the traditional peer-review process. Altogether, this technological ecosystem enables researchers to address more complex questions about human behavior and to do so with greater speed and effectiveness than ever before.

But it's important to remember that these tools, while powerful, are just that: tools. Their value, like that of a hammer, tractor, scalpel, or fermenter, comes from the ability of a knowledgeable professional putting them to use toward a thoughtfully chosen goal. Even the most sophisticated technology cannot compensate for poor research design or flawed methodology.

Thus, while you will have many opportunities to practice with the tools described above throughout the rest of this book, it is important that you also learn the fundamentals of good research design. Understanding how to build a descriptive, correlational, or experimental research project to answer a question you are curious about is a skill that will serve you well. That is why each of the upcoming chapters contains several opportunities for you to develop your knowledge and skill as a behavioral researcher.

Chapter 3: Descriptive Research

If you lived in ancient Babylon, you would have spent more time staring at the stars than you do today. As night fell, there were no lights, no television or phone for entertainment, not even a book to read by candlelight. So, what did people do? They looked up.

Even for the average Babylonian, the heavens were significant. The calendar and many social events were based on the timing of the moon. But for a small group of scholars known as the scribes of Enūma Anu Enlil, the stars were special. Each night these scholars climbed to the top of their temples, armed with sticks for measuring angles, water clocks for tracking time, and clay tablets for writing, and they did something mundane and miraculous: they watched and recorded.

Night after night, year after year, generation after generation, these early astronomers noted the positions of stars and planets, the timing of eclipses, the appearance of comets, and the changing phases of the moon. They used their eyes and instruments to describe the heavens. Yet what made their work truly extraordinary wasn't just their observations, but their painstaking record-keeping. Each night's events were pressed into soft clay tablets, then baked hard in the sun to create permanent records. Over centuries, they amassed thousands of tablets—one of humanity's first scientific databases.

At first, these records were dull—just endless lists of events. "Venus appeared in the east this morning," one might read. "Jupiter stands in the position of the Fish constellation," said another. But with time patterns emerged. The Babylonians noticed that certain events repeated in predictable cycles. They discovered, for example, that Venus appeared as the morning star for 263 days, disappeared for 8 days, then reappeared in the evening for 263 more days. Understanding these patterns allowed the scholars to make predictions.

When would the next lunar eclipse occur? When would Jupiter return to a particular constellation? The records held the answers. And what started as an attempt to describe the stars

eventually led to sophisticated theories about the entire cosmos by astronomers like Ptolemy, Copernicus, and Kepler.

What can the ancient Babylonians teach people today about describing human behavior? One lesson is that before we can predict events or explain why they happen, we must first carefully observe and describe them. This principle—that systematic description is the foundation of scientific understanding—remains just as true today as it was 3,000 years ago.

In this chapter, you will learn how modern behavioral scientists conduct descriptive research. In Module 3.1, we will return to the concept of measurement, introduced in Chapter 1. We will explain how behavioral scientists turn abstract ideas, called theoretical constructs, into numbers that can be analyzed. After discussing measurement, Module 3.2 will explore the methods behavioral scientists use to conduct descriptive research. In Module 3.3, we will return to the survey tools introduced in Chapter 2 and show you how to program a descriptive study. In Module 3.4, you will analyze the data from this study with basic descriptive statistics. Finally, in Module 3.5, you will create your own descriptive question, gather data, and analyze it to describe what people think about a moral dilemma. Throughout the chapter, remember the lesson of the Babylonians: careful description often reveals patterns that lead to predictions and, eventually, to deeper understanding.

Chapter 3 Outline: Descriptive Research

Module 3.1: Basics of Measurement



Reading

Examine the principles of scientific measurement

- Explore different kinds of variables in behavioral research
- Learn about variables in descriptive, correlational, and experimental research

Module 3.2: The Power of Description



Reading

Examine real-world examples of descriptive research to understand how researchers observe and summarize behavior.

- Learn what descriptive research is and how it is conducted
- Review real-world examples of descriptive research

Module 3.3: Designing a Descriptive Survey



Laboratory

Learn to build a descriptive study in Qualtrics

- Read about the Heinz Dilemma used to study morality
- Program a Qualtrics survey to study moral decision making
- Read about how to collect data for this descriptive study

Module 3.4: Analyzing Descriptive Data



Laboratory

Use data from a moral dilemma study to practice analyzing and interpreting descriptive statistics.

- Download data files from the Open Science Framework
- Learn to conduct and interpret descriptive statistics

Module 3.5: Conducting Your Own Descriptive Research



Laboratory

Apply what you've learned by designing and running your own descriptive research project.

- Create your own descriptive study
- Gather data
- Conduct descriptive statistics

Basics of Measurement

Module 3.1

Examine the principles of scientific measurement

Science depends upon measuring things. In chemistry, measures may be the volume, density, mass, or pH of a substance. In biology, they may be the age, growth rate, diet, or metabolic rate of an organism. In astronomy, scientists often measure light, temperature, distance, and the size of celestial bodies. And, in physics, common measures are time, speed, distance, and force. To understand the world, these measures are important.

Measures are important in the behavioral sciences, too. In psychology, scientists measure things like personality, emotions, and memory, as described in previous chapters. In political science, researchers measure which candidate voters prefer and how much people trust the political system. Economists measure how businesses feel about the economy and how much inflation consumers expect in the future. Sociologists measure patterns of social interaction and changes in community attitudes. Each of these measures helps describe what people think and do.

But measurement is often just the beginning. To truly understand people, researchers must carefully describe what they observe. For instance, a single measurement can tell how anxious a person feels right now. But if researchers systematically gather and organize many measurements, they can describe broader patterns, like how anxiety changes throughout the day, varies across situations, or differs between groups of people. Before doing this, however, the researchers need to be confident in what their measures represent.

What are Variables?

Across the behavioral sciences, researchers conduct studies to investigate people's thoughts, feelings, and behavior. Each study measures what are called **variables**. As you learned in Chapter 1, variables are any psychological, physical, or social characteristic that can be assigned a numerical value. When you took the TIPI, it measured the Big Five traits—openness to experience, conscientiousness, extraversion, neuroticism (or emotional stability), and agreeableness (see Box 1.1). Each of these traits is a variable.

Variables, by definition, change from person to person. For example, age is a variable. If you walk down a street in your city or town and ask each person you pass about their age, you will receive answers that vary. Some people may be born in the same year and occasionally you might find people born on the same day, but overall, you will receive answers that differ because age is a characteristic that varies across people. If you were walking with questionnaires that measured personality, household income, or life satisfaction, the answers you receive would vary on these measures, too. For a characteristic to be a variable it must be capable of taking different values across people, situations, or points in time.

Variables are essential building blocks in research. They are what scientists use to form hypotheses, design studies, analyze data, and interpret results. Not all variables are the same, however. Behavioral scientists distinguish between independent variables, dependent variables, and control variables. You will learn about the differences between variables in later chapters.

A variable can be contrasted with a **constant**. Whereas variables take different values from person to person, constants stay the same. For example, if everyone in a study was born in the United States, country of birth would be a constant.

Sometimes, researchers intentionally transform a variable into a constant. For example, imagine researchers conducting an experiment to see if feeling afraid makes people seek social connection with others. The researchers might be concerned that men and women will respond differently because of social expectations, as men are often expected to hide their fear whereas women may express it more openly. To make their study simpler, the researchers might include only women or only men (e.g., Schacter, 1959). Doing so would make gender a constant, allowing the researchers to focus specifically on how fear affects social connection without gender differences complicating the results. In Chapters 5 and 6 you will learn more about how variables and constants can be combined within a study to understand the causes of behavior.

Conceptual vs. Operational Definitions

If you look across the behavioral sciences, you will find some variables are tangible and others are abstract. Heart rate, for instance, is tangible—it can be felt, heard, and measured with various devices. Household income is also tangible; there are dollars that represent the money flowing

into a household during the year. Life satisfaction, on the other hand, is abstract. There is nothing behavioral scientists can point to and say, 'That is life satisfaction,' although sipping a Mai Tai on the beach in Hawaii might be close!

Abstract variables in the behavioral sciences are called **theoretical constructs** because they are difficult to observe directly. The challenge for behavioral scientists lies in taking a theoretical construct like anxiety, extraversion, happiness, or regret and finding a way to measure it. This process is both challenging and important because it's only through measurement that a construct can become a variable.

To conduct a scientific study, a researcher must define both the variable they are measuring (the concept) and how they are measuring it (the operational definition). An operational definition specifies exactly how the researcher will turn the concept into a quantitative measure, where each person is assigned a number that corresponds to the concept being measured.

Consider anxiety. Everyone knows what it feels like to worry. But how can a psychologist translate the subjective experience of anxiety into a precise measure where each person who is studied receives a score?

A common method is to ask questions. These questions often form a scale or **measurement instrument**. To measure anxiety, for example, many behavioral scientists use the Generalized Anxiety Disorder 7-item scale, abbreviated GAD-7 (Spitzer et al., 2006). The items from this instrument are in Table 3.1.

By asking people to complete the GAD-7, researchers can translate each person's feelings of anxiety into a number. By summing the numbers from all items within the scale, researchers obtain an overall score that can range from zero (no anxiety) to 21 (severe anxiety). This score represents a measurement of anxiety and captures differences between people. In other words, using the GAD-7, researchers can measure anxiety and make it into a variable.

Item #	Problem	Not at All (0)	Several Days (1)	More than Half the Days (2)	Nearly Every Day (3)
1	Feeling nervous, anxious, or on edge				
2	Not being able to sleep or control worry				
3	Worrying too much about different things				
4	Trouble relaxing				
5	Being so restless that it is hard to sit still				
6	Becoming easily annoyed or irritable				
7	Feeling afraid as if something awful might happen				

Table 3.1. The GAD-7 is an operational definition of anxiety. It relies on people's responses to seven questions that assess anxious feelings over the previous two weeks.

Psychometrically, there is nothing special about the GAD-7 or anxiety. Behavioral scientists have developed several ways to measure anxiety (e.g., Newton & Buck, 2000; Nomura et al., 2006; Shelton et al., 2010; Vigil-Colet et al., 2008) and thousands of ways to measure other psychological constructs. Each measure follows the same process: take a behavior or experience, create an instrument to assess it, and assign each person a number. These numbers are the basis of behavioral research. In the next chapter, you will learn how to create measurement instruments for virtually any construct you would like to measure.

Three Types of Research: Descriptive, Correlational, and Experimental

After behavioral scientists develop a way to measure the variables they are interested in, they put those measures to work in research studies. There are three kinds of studies in the behavioral sciences: those that *describe* behavior, those that *demonstrate a relationship* between variables, and those that *establish cause and effect* relationships. To understand how measurement moves scientists toward each objective, consider the following example.

When people run a marathon—a race that lasts 26.2 miles—their finishing times vary. Figure 3.1 (left panel) shows how much they vary (Allen et al., 2016). There are three things to note about this figure.

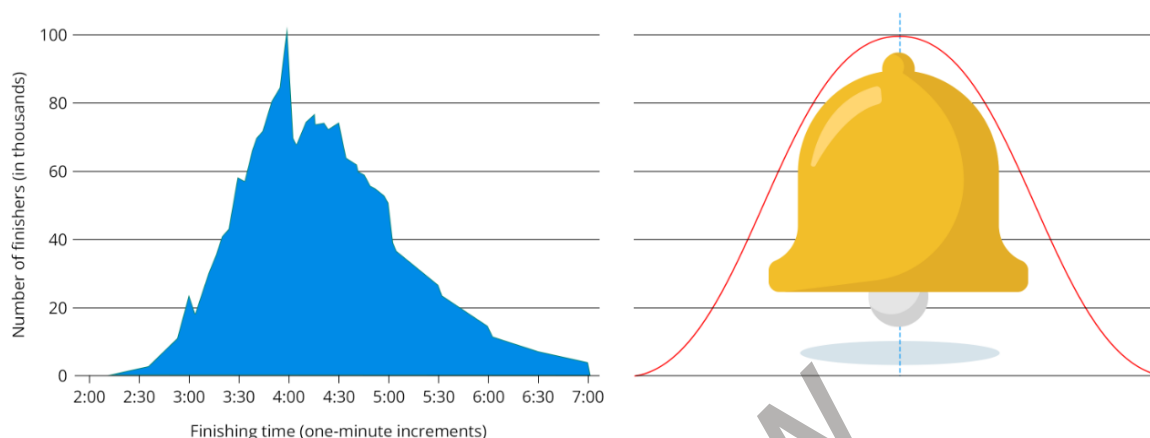


Figure 3.1. Left: Marathon finishing time for almost 10 million runners. Source: Allen et al., 2016. Right: The outline of a bell is where people derive the term ‘bell curve’ from, which is used to describe a normal distribution of data.

First, there is A LOT of variation between people. The fastest runners finished in under 2 hours and 30 minutes—less than six minutes per mile!—while the slowest runners took 7 hours, or about 16 minutes per mile. Second, most people are in the middle. The highest line in the figure is at the four-hour mark, with more than 100,000 people achieving this time. The bulk of the other times are between 3 hours and 5 hours. Finally, people’s finishing times are close to a normal distribution. A normally distributed variable is one in which most values cluster around the middle and fewer values appear as you move toward the tails, or sides, of the figure. Normal distributions are often referred to as a ‘bell curve’ (see Figure 3.1, right panel).

Now, imagine a researcher is interested in the psychology of sport and performance. With a large dataset like the one above, the researcher can describe a lot about people running marathons. Thanks to the richness of the description, the researcher is likely to wonder what explains the differences between runners. In other words, why do some people finish faster than others? The search for that answer moves into the realm of correlational and experimental research.



Stop and Think!

Before reading further, take a minute to consider which variables might predict how quickly someone runs a marathon. These predictor variables could be physiological, psychological, or social. Write down your ideas and then see how they align with the variables mentioned below.

When predicting marathon times, of course, physical characteristics matter. Some people are blessed with a heart and lungs that efficiently move oxygen-rich blood around the body, and these people tend to run fast. Some people have great biomechanics—things such as a long stride, a strong cadence (steps per minute), or even how the foot strikes the ground—that make them fast runners. A researcher interested in predicting marathon times might consider these physiological variables.

But many psychological variables may predict marathon times just as well. These include things like mental toughness, pain tolerance, goal setting, stress management, previous race experience, and even strategies for managing boredom. Each of these psychological variables is something a behavioral scientist could measure and use to predict differences in marathon times. After establishing an association between two variables, the researcher might test whether one variable causes the other by conducting an experiment. But before researchers can test predictions or uncover causes, they first need to know what's out there—and that's where descriptive research begins.

The Power of Description

Module 3.2

Examine real-world examples of descriptive research to understand how researchers observe and summarize behavior.

What is Descriptive Research?

When researchers want to describe what people do, think, or feel, they conduct descriptive research.

Describing behavior and mental processes can be valuable. That is because description is often the first step in creating theories that explain when and why certain behaviors occur or in creating interventions intended to change behavior. For example, if you wanted to know how common different types of cancer are, how much credit card debt the typical household carries, how many people purchase Tide laundry detergent, or how fast people run a marathon, you would need to conduct descriptive research.

A Real-World Example: Describing Mental Health Among U.S. Adults.

Within the United States, the National Institutes of Mental Health has a mandate to promote awareness of mental illness and to guide policy for treatment. To fulfill this mandate, the agency needs to know how many people suffer from mental illness—a question for descriptive research.

During 2021, the National Institutes of Mental Health conducted a large survey to understand what percentage of U.S. adults experienced some form of mental illness throughout the year. To conduct this study, the researchers compiled a list of every residential address in the country (NIMH, 2023). Then, they selected households to participate. The data collection stretched across an entire year and ended with more than 70,000 interviews!

Figure 3.2 shows the results. As seen in the left-most bar, the overall rate of mental illness was 22.8%. Across social groups such as sex, age, and race, the rate of mental illness varied. Women, for instance, were more likely to experience mental illness than men, younger people were more likely to experience mental illness than older people, and White people were more likely to experience mental illness than African American, Asian, or Hispanic people.

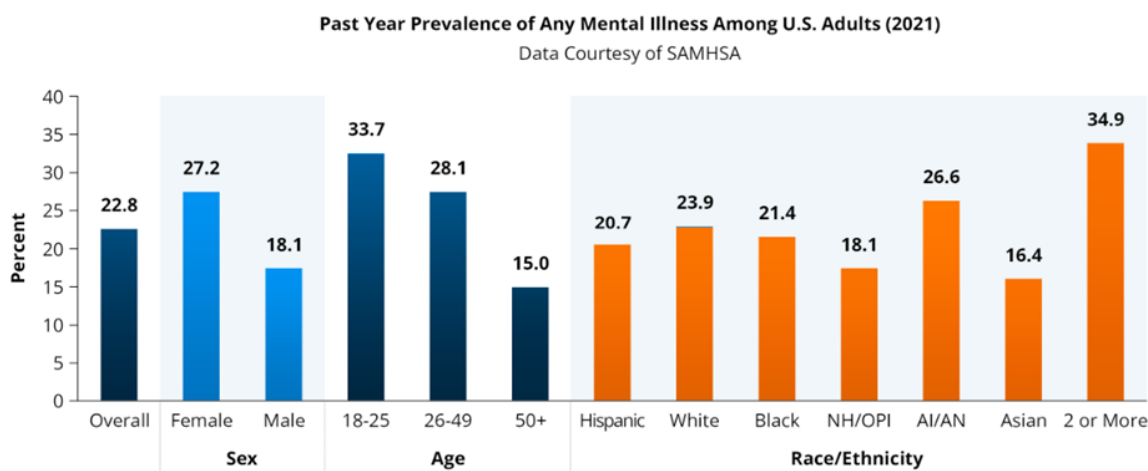


Figure 3.2. The prevalence of mental illness among U.S. adults in 2021. Data are from the National Institutes of Mental Health annual National Survey on Drug Use and Health (NSDUH).

The descriptive data gathered by the National Institutes of Mental Health have important consequences. Government agencies from the White House all the way down to state and local health departments use the data to allocate funding and to craft policies aimed at treating mental illness (National Survey on Drug Use and Health, n.d.). University researchers use the data to create studies that seek to understand why rates of mental illness may be increasing or decreasing within certain groups over time. And clinicians use the data to create educational materials and outreach campaigns aimed at providing proper treatment for mental health issues. In other words, the data from this one descriptive study inform important decisions that guide future research and interventions intended to treat mental health.

Gathering Descriptive Data with Behavioral Sensors.

While surveys are a common way to gather descriptive data, they are not the only way. As described in the previous chapter, researchers often use different tools to observe behavior.

To illustrate what observational methods can look like, consider a study in which researchers wanted to know how much water people use while showering (Pereira-Doel et al., 2024). Water use is a growing concern in many countries because people are quickly depleting the supply of groundwater (Rojanasakul et al., 2023). Heating water for showers also consumes energy that is thought to contribute to climate change. Before proposing solutions to either of these problems,

however, the researchers wanted to understand the problem by describing people's showering habits.

To do that, they installed sensors—a device that passively records input from the physical environment—in hundreds of showers around a university campus. The sensors recorded how long people showered over 39 weeks, resulting in 86,000 observations (the sensors only recorded water usage to protect people's privacy). From this data, the researchers calculated that the average shower lasted 6.7 minutes (Figure 3.3). The sensors determined the water pressure of each shower and how much water was used. From these descriptive statistics, the researchers developed an understanding of how much water was used in showers and a plan for future research aimed at reducing water use. This is exactly how descriptive research is supposed to work.

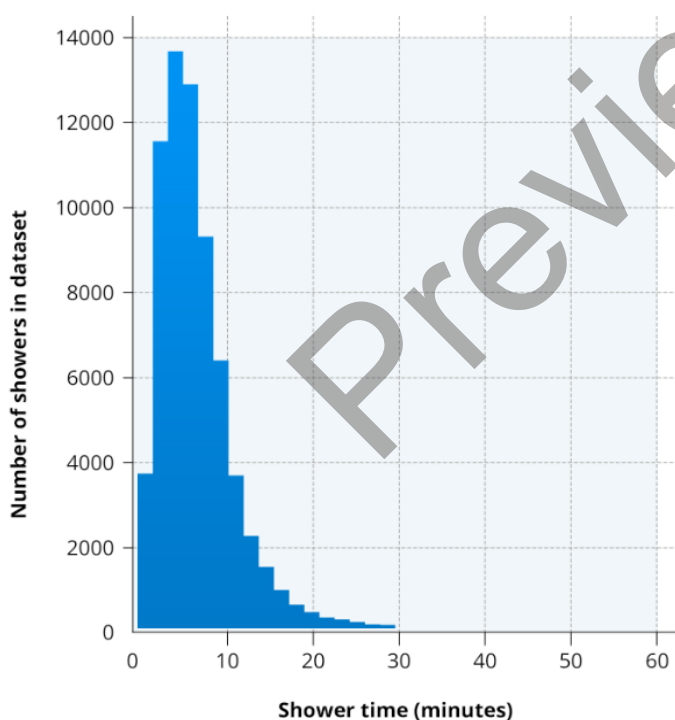


Figure 3.3. This figure shows the duration of showers detected by smart sensors installed on a university campus (from Pereira-Doel et al., 2024).

Designing a Descriptive Survey

Module 3.3

Learn to build a descriptive study in Qualtrics

So far, we have explored how researchers measure variables and use those measurements to describe human behavior. We have seen examples ranging from marathon times to the prevalence of mental illness and how long people shower. While these examples illustrate different aspects of descriptive research, they do not show you how a study comes together. To understand that let's walk through an example from start to finish.

In this section, we will examine a descriptive research project that investigated how people respond to a moral dilemma. You will see how researchers designed questions to measure abstract concepts, how they implemented those questions on a survey platform, how they gathered data from online participants, and how they analyzed the results to describe people's thoughts and behavior. Afterward, you will apply what you've learned to conduct your own descriptive project.

Describing People's Reactions to a Moral Dilemma

In the 1950s, psychologist Lawrence Kohlberg began studying how people reason about right and wrong. To conduct his studies, Kohlberg presented people with moral dilemmas that pit different moral principles against each other. His most famous scenario, known as the Heinz dilemma, asks people to weigh the sanctity of human life against the law.

Here is a modern version of the dilemma:

A woman is on her deathbed, and the only drug that doctors say could save her is a newly discovered form of medication. This drug is expensive to produce. The company that manufactures the drug also sells it at a significant markup. Each dose of the medication costs the company \$20,000 to produce, but they sell it for more than \$200,000 per dose.

The woman's husband, Heinz, exhausts all his options trying to gather the money. He asks family and friends, he starts a GoFundMe, he takes a second job, and he appeals to the company for a discount. Altogether, he only manages to collect about \$100,000—half of what the drug costs.

Heinz pleads with his insurance and the company that makes the drug, explaining his wife's dire condition and asking for a discount or a payment plan, but everyone refuses. Faced with the imminent death of his wife and unable to afford the drug legitimately, Heinz feels compelled to break into a pharmacy and steal the medication.



Stop and Discuss!

Before we explore how researchers study people's reactions to the Heinz dilemma, take a moment to think about your own response. What would you do in Heinz's shoes? Here are some questions to consider. Talk these over with your friends or classmates. You will be surprised at the different opinions you encounter.

1. What do you think Heinz should do? What was your initial reaction to the dilemma, and did your thinking change as you considered it more deeply?
2. What principles or values influenced your decision? For instance, how do you weigh the value of human life against respect for property rights and laws?
3. How might different people or cultures view this situation? What factors could influence whether someone thinks Heinz's actions were justified?
4. If you were designing a study to understand how people view this dilemma, what questions would you ask? What aspects of people's responses would be most interesting to measure?
5. Kohlberg used this dilemma to study how moral reasoning develops as people age. He conducted lengthy interviews asking not just whether people thought Heinz should steal the drug, but why they reached their conclusion. During these interviews, participants would discuss their thought process, describe their values, and explain how they weighed different moral principles against each other. Kohlberg then carefully reviewed the transcripts of these interviews, looking for patterns in how people reasoned about morality.

6. The process of conducting in-depth interviews and systematically analyzing people's responses to identify themes is called **qualitative research**. Qualitative research does not follow the process of operationalizing variables with numbers. Instead, it involves carefully reading open-ended interviews and extracting patterns and ideas from the conversations. Qualitative research is an important method in many fields of behavioral science, and we will examine it more in Chapter 8.
7. Through qualitative analysis, Kohlberg identified six distinct stages in how people think about moral problems. In this study, however, we will focus on describing quantitative outcomes.

Choosing What to Measure

When conducting descriptive research, behavioral scientists must decide exactly how to measure the concepts they are interested in. For the Heinz dilemma, we wanted to understand two aspects of people's reactions: what they thought Heinz should do, and how morally acceptable they found his actions.

For the first aspect, we chose a simple yes/no question: "Should Heinz have broken into the pharmacy to steal the drug for his wife?" This question forces people to decide, just as Heinz had to do.

However, we recognized that people's views might be more nuanced than a yes or no response. Someone might think Heinz shouldn't steal the drug but still see his action as somewhat justified, given the circumstances. So, we added a second question asking: "How morally acceptable was it for Heinz to steal the drug?" Participants answered using a 7-point scale ranging from "Not at all acceptable" to "Entirely acceptable."

Once the measures for a study have been selected, the next step is to program them into a survey. In the activity that follows, we will show you how to program the Heinz dilemma in Qualtrics.

Research Activity 3.1: Building the Heinz Dilemma in Qualtrics

Let's look at how to create this study in Qualtrics. Setting up the study requires six simple steps. These steps are described below and demonstrated in the accompanying video: [enter link].

Step 1: Setting Up Your Project

First, login to your Qualtrics account by using the credentials your instructor provided or creating a free student account: <https://www.qualtrics.com/free-account/>.

After logging in, click the “Create Project” button. Select “Survey” from the options provided, then choose “Blank Project.” Name your survey “Heinz Dilemma Descriptive Project” and click “Create Project.”

You will now see the Qualtrics Survey Builder (Figure 3.4). Notice the menu on the left side showing icons for the “Builder” “Survey Flow,” “Look and Feel” and other options. Also notice the tabs across the top for things like “Distributions,” “Data & Analysis,” and “Reports.” For now, we will build the questionnaire on the survey tab.

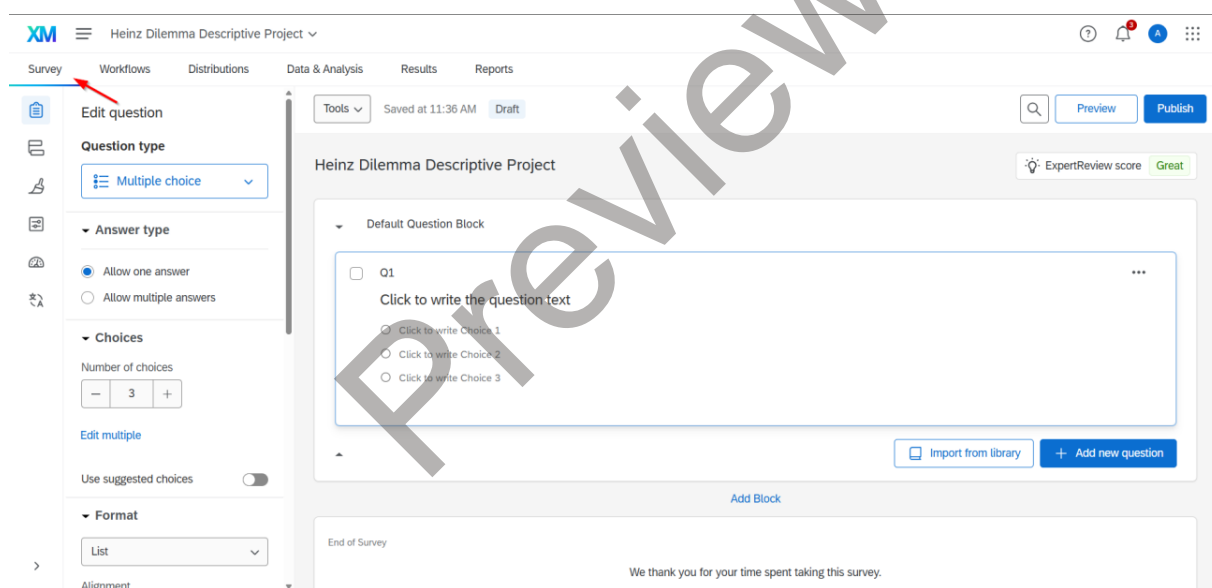


Figure 3.4. A view of the Survey Builder in Qualtrics.

Step 2: Organizing the Survey into Blocks

Professional researchers organize their surveys into logical sections called “blocks.” For this study, we will create three blocks:

1. Introduction and Consent

2. Heinz Dilemma
3. Demographic Questions

Look for the “Default Question Block” that appears in your survey. Click on the block title to rename it “Introduction and Consent.” Then click the “Add Block” button at the bottom to create two more blocks. Name them “Heinz Dilemma” and “Demographic Questions.” The result should look like Figure 3.5.

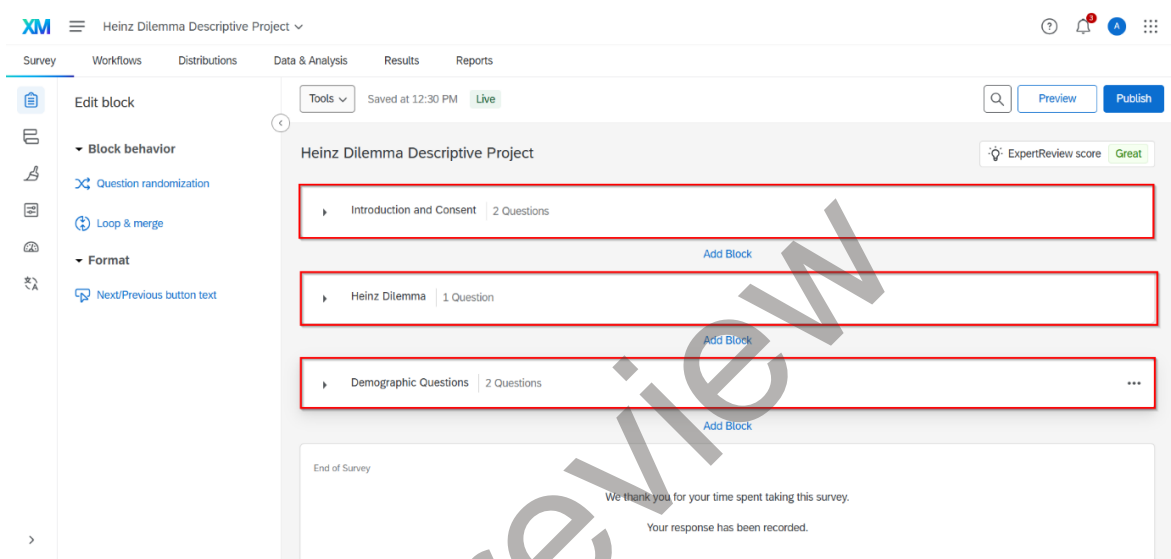


Figure 3.5. A picture of three blocks within the survey builder

Step 3: Creating the Introduction

In the Introduction block, click “Add Question.” Choose “Text/Graphic” from the question types. In the box that appears for editing, enter the text below to introduce your study:

“Welcome to the Perceptions of Social Events Study. This study examines how people respond to a social situation. Your participation will involve reading a scenario and answering a few simple questions. It will take approximately 3 minutes to complete. Your responses will be anonymous and used for research purposes only. Please make sure to complete the study in one sitting and in a quiet location without distraction.

By clicking ‘Next’ below, you indicate that you are at least 18 years old and voluntarily agree to

participate in this study. You may exit the survey at any time if you wish to discontinue participation.”

Add another question and choose multiple choice. Using the menu on the left, limit the answer options to two and paste the content below into the question text:

“By clicking ‘Continue’ below, you indicate that you are at least 18 years old and voluntarily agree to participate in this study. You may exit the survey at any time if you wish to discontinue participation.”

Finally, edit the two answer options to read: 1) “Continue with the study”, and 2) “Exit the study.”

Within the survey, this information will look like Figure 3.6. If you were actually gathering the data for this project, the consent form would be longer (if mandated by your institution; see Chapter 15) and you would need to ensure people who select “End participation” are taken to the end of the survey. We will ignore these details for now.

Welcome to the Perceptions of Social Events Study. This study examines how people respond to a social situation. Your participation will involve reading a scenario and answering a few simple questions. It will take approximately 3 minutes to complete. Your responses will be anonymous and used for research purposes only. Please make sure to complete the study in one sitting and in a quiet location without distraction.

By clicking 'Continue' below, you indicate that you are at least 18 years old and voluntarily agree to participate in this study. You may exit the survey at any time if you wish to discontinue participation.

Do you wish to participate in this study? Click Continue to proceed or Exit to end the study.



Continue with the study

Exit the study

Figure 3.6. The first thing participants see in a study is often welcome information and a question asking if they consent to participate.

Step 4: Adding the Dilemma and Outcome Measures

In the Heinz Dilemma block, add the dilemma participants will read. Click “Add Question” and select “Text/Graphic” from the question types. Then paste the Heinz dilemma from earlier in the chapter into the question box. You may need to create spaces between paragraphs to ensure the dilemma is easy for participants to read.

The result should look like Figure 3.7 in the survey builder. When the survey is presented to participants it will appear as in Figure 3.8.

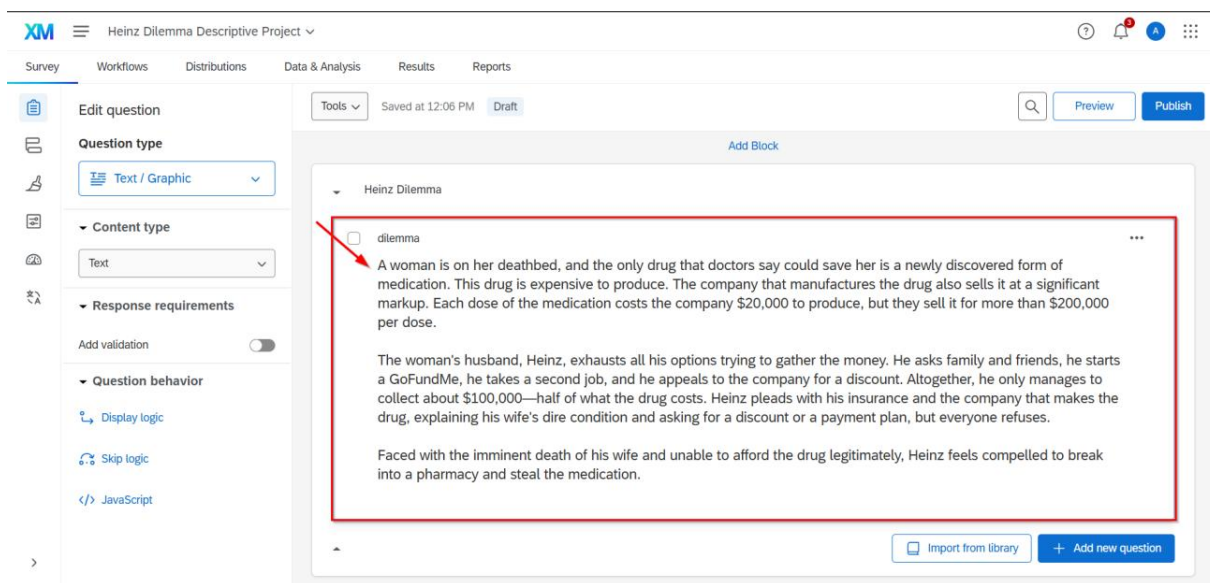


Figure 3.7. The Heinz dilemma within the survey builder.

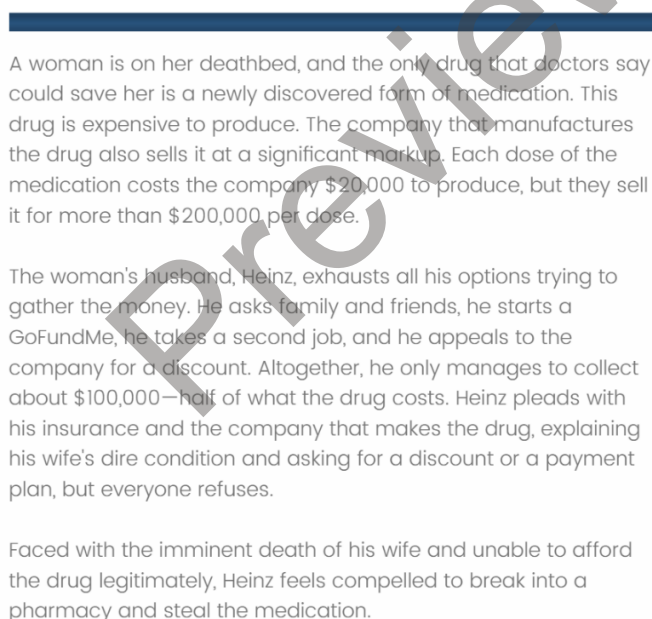


Figure 3.8. The Heinz dilemma as it appears to participants.

After the dilemma, enter a “page break.” Page breaks divide the content within a block so that participants see each question on a separate screen. After the page break, add a multiple-choice question. For the question text write: “Should Heinz have broken into the pharmacy to steal the

drug for his wife?” Then, edit the number of answer options down to two, and write “yes” and “no” into the answer options. The result should look like Figure 3.9.

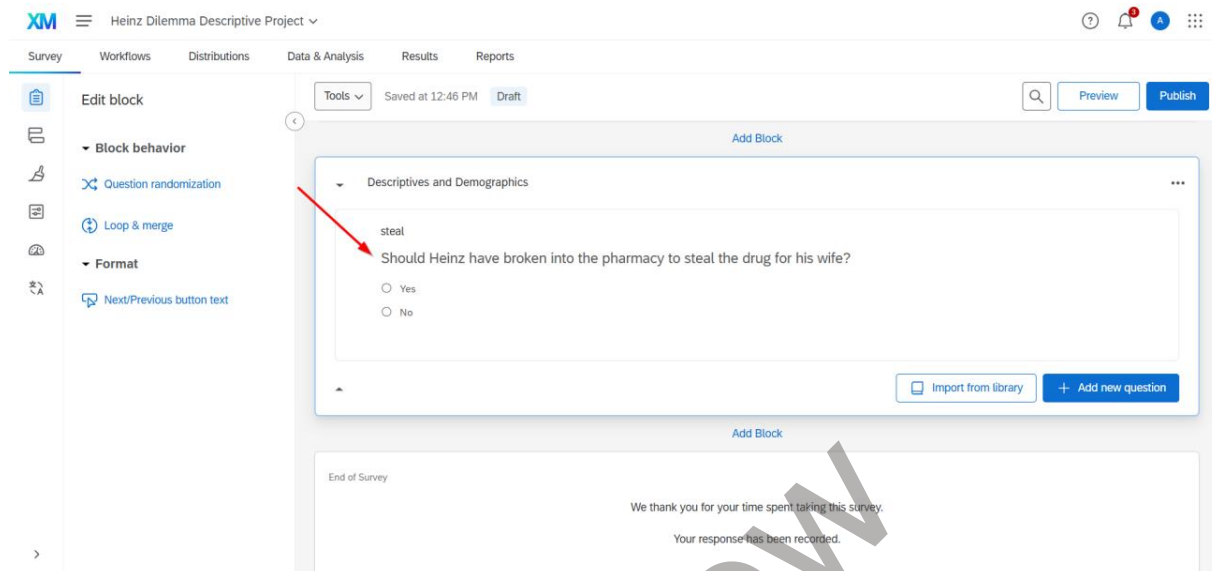


Figure 3.9. A question about whether Heinz should have stolen the drug in the last block of the survey.

After the yes/no item, add another multiple-choice question. Paste the question, “How morally acceptable was it for Heinz to steal the drug?” into the question text. Then, edit the answer options so that there are 7 choices ranging from 1 “Not at all acceptable” to 7 “Entirely acceptable.” The result should look like Figure 3.10.

Now, you are ready to move on to the last block of the survey.

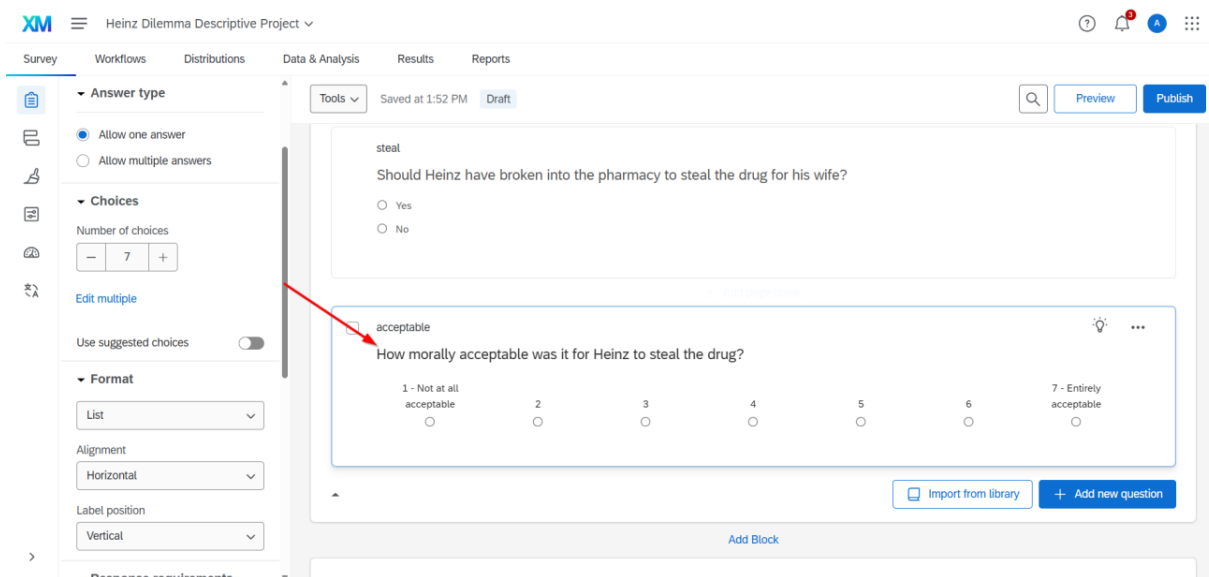


Figure 3.10. A question asking people how acceptable it was for Heinz to steal the drug.

Step 5: Adding Demographic Questions

In the Demographic questions block, you can add questions that gather information about who is participating in the study. These include things like people’s age, gender, race, and education level.

The easiest way to add these questions is with Qualtrics’s question library. This library will not be appropriate for every study, but for this study you can click the “Import from library” button within the demographics block. In the side window that appears, select “Demographics” under the Qualtrics certified header, and then choose “Global demographics.” The first question should be about age. If you click on the item, you will see a preview of how it appears and a button to “Insert question” (Figure 3.11). Add the question.

If you explore the Global and U.S. Demographics tabs, you should find questions asking about participant’s gender, race, education level, and household income. Add each of these to the demographics block, and then the survey is done.

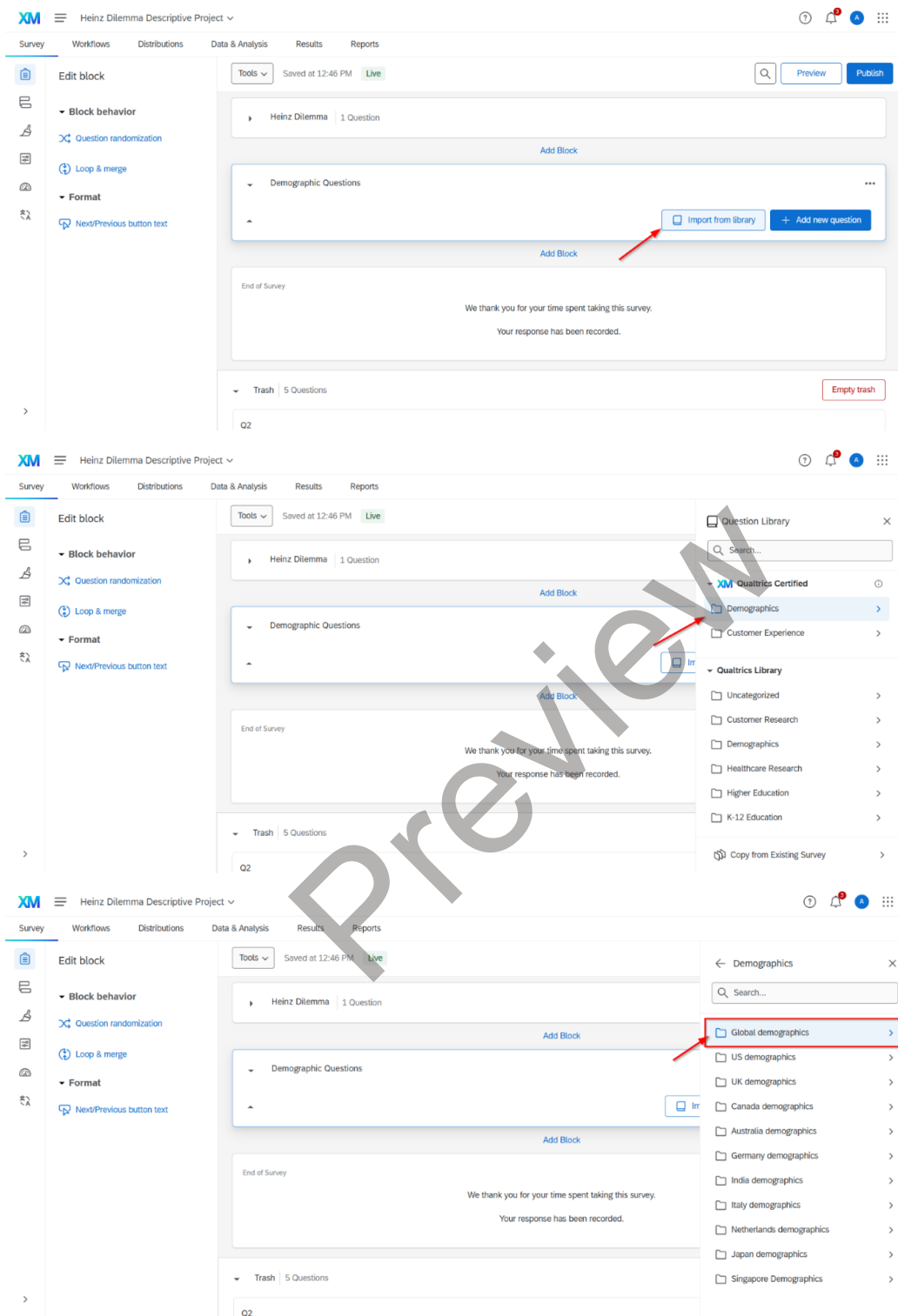


Figure 3.11. A) The top photo shows where to import questions from a library, B) The middle photo shows where to select Qualtrics Certified demographic questions, C) The bottom photo shows where to select Global Demographic questions.

Step 6: Previewing Your Survey and Generating the Survey Link

Once you have built the survey, you should preview it. Previewing allows you to see how the survey appears to participants and to make sure that everything looks professional and works as expected.

To preview, click on the “Distributions” tab in the menu. On the following page, select the option for “Anonymous link.” You should see something that looks like Figure 3.12. You can copy this link into a new browser to preview the survey. Later, you can also use this same link to send the survey to participants or paste it into a participant recruitment site like Connect.

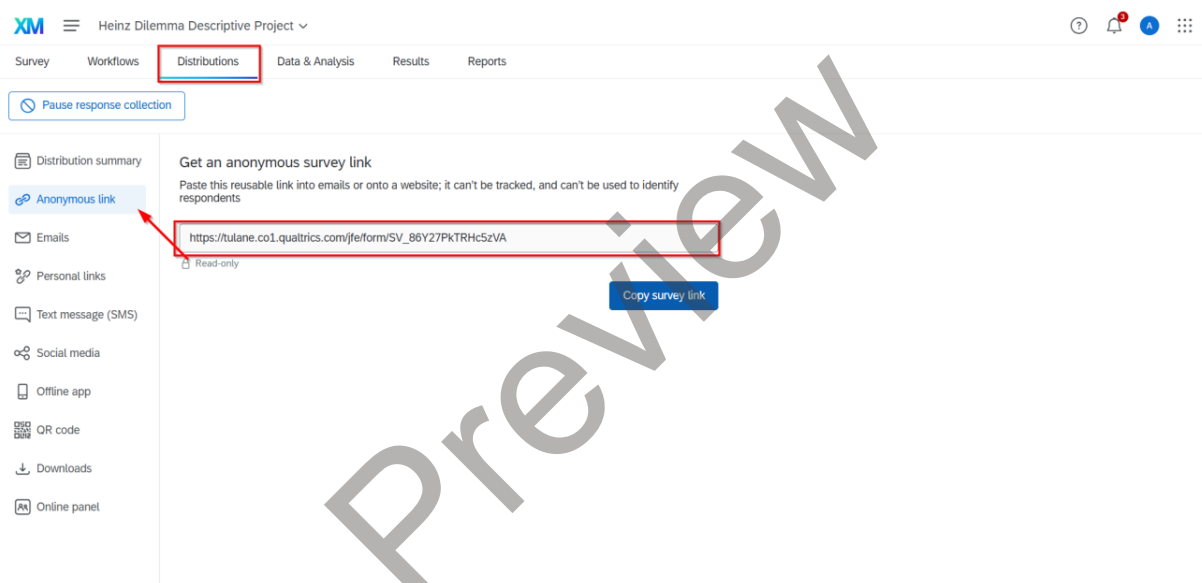


Figure 3.12. An “anonymous link” can be sent to participants or pasted into a participant recruitment site. Anyone with the link can access your survey.

Congratulations! You have just created a survey that can be used to describe people’s reactions to the Heinz dilemma. In creating this survey, you have seen how researchers use survey platforms like Qualtrics to organize studies into logical blocks, add stimuli like the moral dilemma, and program different types of questions (text, matrix tables, multiple choice). You have also seen how a survey platform that looks a bit technical on the back end, can render clean and professional-looking surveys that guide participants through a research experience.

Gathering Descriptive Data Matched to the U.S. Census

Normally, the next step after programming a survey is to gather data. We have done that for you.

We gathered data from 100 participants on Connect. Before launching the survey, we had to consider several practical details like how long the study would take and how much to pay participants. Through pilot testing, we found that most people completed the survey in 2-3 minutes. Following standard payment practices, we offered 50 cents for taking the study—equivalent to about \$12 per hour.

In addition to payment, we had to consider *who* to sample—an important decision in any descriptive research project. We wanted to describe how people in the U.S. react to the Heinz dilemma, so we needed participants who represented the demographics of the U.S. population. Connect offers a feature called Census Match, which creates quotas that helped achieve this goal. The Census Match quotas specified what percentage of participants should come from different demographic groups.

We created **quotas** to match the U.S. Census numbers for age, gender, race, and ethnicity. For example, about 13% of the U.S. population identifies as Black or African American. Using the quotas, we ensured 13% of our sample did too. Quotas matched to the U.S. Census resulted in a sample that was reasonably matched to key demographics of the U.S. population (Chapter 9 discusses representativeness in detail). Within ten minutes of launching the study, we had responses from 100 participants. In the next module, we will show you how to analyze this data, before giving you the chance to collect your own descriptive data at the end of the chapter.

Analyzing Descriptive Data

Module 3.4

Use data from a moral dilemma study to practice analyzing and interpreting descriptive statistics.

Downloading the Heinz Dilemma Data From OSF

As mentioned in Chapter 2, the materials for each project in this book are stored on the *Research in the Cloud* OSF page. You can access that page here: <https://osf.io/a8kev/>.

When you arrive on the OSF page, click the “Files” option in the menu (Figure 3.13). Then, choose the folder labeled “Ch. 3 – Descriptive Research.” Inside, download the file named “RITC_DATA_CH03_HeinzDilemma.sav” (Figure 3.14). All files in this book follow this naming convention. RITC stands for *Research in the Cloud*—the name of this book. Then, the first word describes the file type—DATA, SURVEY, SYNTAX, or MATERIALS—followed by the chapter number and a description of the project.

The screenshot shows the OSFHOME interface for the project "Research in the Cloud: A Guide to Online Behavioral Science". The navigation menu includes "Files", "Wiki", "Components", "Analytics", and "Registrations". The "Files" tab is selected and highlighted with a red arrow. Below the navigation, the project title is displayed, along with contributors (Aaron J Moss, Rachel Hartman, Leib Litman), creation date (2024-04-10 11:11 AM), and last update date (2025-03-20 04:15 PM). The "Files" section is currently empty, with a message: "Files can now be accessed and managed under the Files tab." The "Recent Activity" section shows three entries, all dated 2025-03-20 04:15 PM, indicating file renames by Aaron J Moss.

Figure 3.13. The “Files” tab will take you to folders with files you can download for each project.

If you are using SPSS, download the .sav data file; if you're using another statistical program, choose the .csv file. Once the file is open in your analysis program, you are ready to analyze the data.

The screenshot shows the OSF Storage interface for a project titled 'Research in the Cloud: A Guide to Online Behavioral Science'. The left sidebar contains navigation options: Overview, Metadata, Files (selected), OSF Storage, Wiki, Components, Analytics, and Registrations. The main content area displays the folder 'Ch. 3 - Descriptive Research' with a list of files. A red arrow points to the file 'RITC_DATA_CH03_HeinzDilemma.sav', which is highlighted in blue. The file list includes the following information:

File Name	Downloads	Size	Timestamp
RITC_DATA_CH03_HeinzDilemma.csv	0 Downloads	21.6 kB	2025-04-21 10:17 AM
RITC_DATA_CH03_HeinzDilemma.sav	0 Downloads	364.1 kB	2025-04-21 10:17 AM
RITC_MATERIALS_CH03_HeinzDilemma.docx	0 Downloads	30.5 kB	2025-04-21 10:19 AM
RITC_SURVEY_CH03_HeinzDilemma.qsf	0 Downloads	30.2 kB	2025-04-21 10:13 AM

At the bottom of the interface, there is a footer with copyright information: 'Copyright © 2011-2025 Center for Open Science | Terms of Use | Privacy Policy | Status | API TOP Guidelines | Reproducibility Project: Psychology | Reproducibility Project: Cancer Biology' and a 'Help' button.

Figure 3.14. Within the folder for each chapter you will find files you can download.

Research Activity 3.2: Analyzing People's Response to the Heinz Dilemma

In Chapter 2, we introduced SPSS as a common statistical package within the behavioral sciences. Now, you have your first chance to use it.

The video we created for this assignment gives an overview of SPSS and shows you how to perform the descriptive analyses: [link to video]. The instructions in Box 3.1 also provide an overview of the analyses. Let's look at the data.

HOW TO: Describe People's Responses to the Heinz Dilemma

Follow these steps to obtain descriptive statistics for the question about whether Heinz should have stolen the drug and how acceptable his behavior was

Open the dataset

- Open SPSS and navigate to File --> Open --> Data
- Find the "RITC_DATA_CH03_HeinzDilemma.sav" file from where you downloaded it

Frequency analysis for the yes/no question

- Click "Analyze" in the top menu
- Select "Descriptive Statistics --> Frequencies"
- Find the variable named "Steal" in the left panel and move it to the Variable(s) box
- Click "Charts" and select "Bar Charts" --> "Percentages" --> Continue
- Click "OK" to run the analysis

Descriptive statistics for the acceptability item (1=7 scale)

- Click "Analyze" in the top menu
- Select "Descriptive Statistics --> Descriptives"
- Move the "Acceptable" variable to the Variable(s) box
- Click "Options" and check the boxes for "Mean," "Std. Deviation," "Minimum" and "Maximum." Then click "Continue" --> "OK"

Box 3.1. Instructions to obtain descriptive statistics for the Heinz dilemma.

For the first measure—whether Heinz should steal the drug—we simply calculated the percentage of participants who answered "yes" versus "no." The results revealed an interesting split: only 47% of participants thought Heinz should steal the drug. This suggests people were divided about Heinz's choice, with a slight majority feeling he should not break the law even to save his wife's life.

However, the results from the second measure showed a different pattern. When people were asked how morally acceptable it was for Heinz to steal the drug on a scale from 1 (not at all acceptable) to 7 (entirely acceptable), they gave an average rating of 4.39 (Figure 3.15). Given that 4 indicates neutrality, the data showed more people felt that Heinz's actions were morally

acceptable. Indeed, the most common response (the mode) was 7, meaning that more people saw Heinz's action as morally justified even though they didn't think he should have stolen the drug.

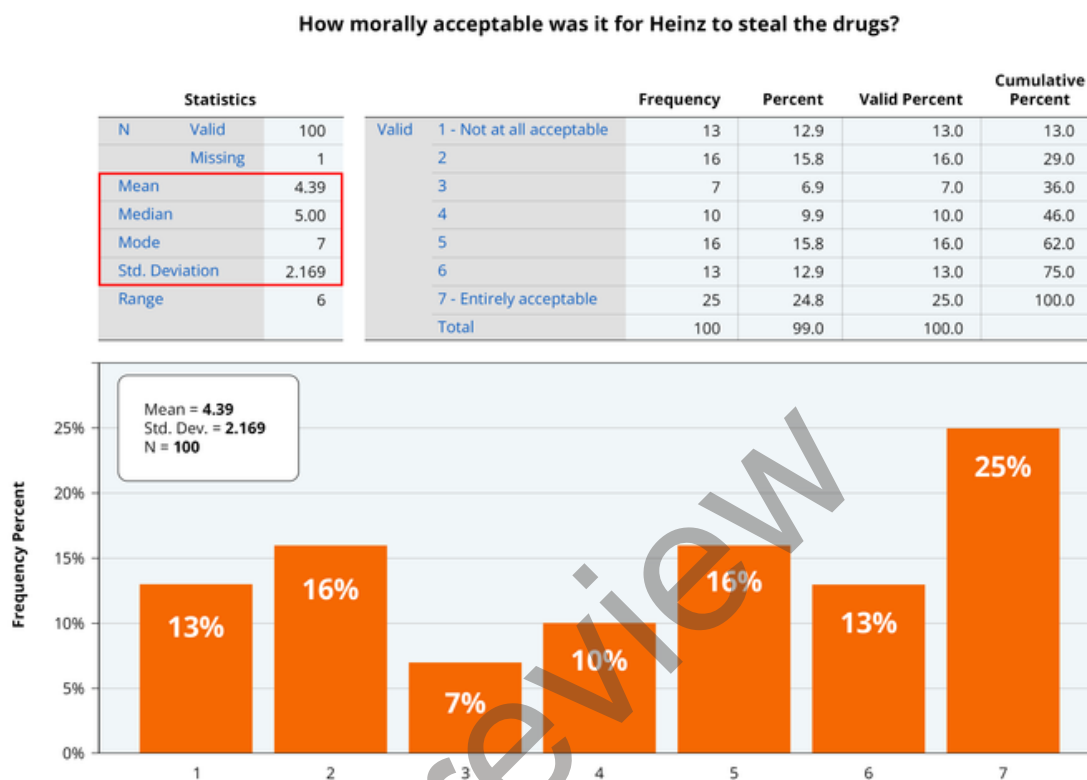


Figure 3.15. Results of the question asking people how acceptable it was for Heinz to steal the drug on a 1 (*not at all acceptable*) to 7 (*entirely acceptable*) point scale.

What Descriptive Data Reveal about Moral Judgments

When people were forced to make a choice (yes or no), most leaned against stealing the drug. But when people judged the morality of the action, many indicated that Heinz's action was morally acceptable. Why the difference?

Although the data from this study cannot say for sure, the pattern suggests many people were influenced not just by sympathy for Heinz's situation but also by a concern for rules and societal order. While participants recognized the importance of saving a life, they ultimately prioritized respect for law and order. This may explain why most people said stealing the drug was wrong, even though many of them found the action morally acceptable, on some level.

People's reasoning in this study seem to align with what Lawrence Kohlberg described as Stage 4 moral reasoning. People at this stage emphasize the importance of laws and social stability. They tend to believe that laws exist for the benefit of society as a whole and that people should adhere to them, even when faced with morally challenging situations. Under this logic, breaking the law, even for a good reason, could set a precedent that weakens society.

Preview

Conducting Your Own Descriptive Project

Module 3.5

Apply what you've learned by designing and running your own descriptive research project.

Research Activity 3.3: Creating Your own Descriptive Study (and putting it in Qualtrics)

Now that you have seen an example of descriptive research, it's time to conduct your own project. We want you to create a version of the Heinz dilemma that explores another aspect of moral reasoning. In fact, you can think of this project as a competition. Here is what you need to do.

First, create a variation of the Heinz dilemma by changing a key element. This can be anything, such as the relationship between the characters, the stakes or consequences of the theft, or even the context of the dilemma. To give you an example, students in one class wanted to know how people's judgments would change if Heinz was stealing the drug to save his dog, instead of his wife.

After you have an idea, write it down and develop a hypothesis. What percentage of people do you think will say Heinz should steal the drug and will that match their ratings of acceptability? How do you think your scenario might compare to the original?

Once you have an idea and a hypothesis, modify the Qualtrics survey. Replace the original dilemma with your new scenario, and consider which questions to keep, modify, or add. Each student should pitch their idea in a short presentation (one minute or less to outline the idea). Then, the class can vote on the best idea and collect the data. Once the data collection is complete, the creator of the survey can share the file with the class, and everyone can analyze it.

To ensure your study is ready if your idea wins, you can set up and save a Connect project as part of this assignment. We guide you through how to do so below.

Gathering Descriptive Data Online

The video we created for this assignment will guide you through creating a study on Connect: [enter link]. To demonstrate some of the steps below, we will use the Heinz Dilemma survey you created earlier in the chapter.

To get started, login to the researcher side of Connect, just as you did in Chapter 2. Navigate to: <https://account.cloudresearch.com/Account/Login>, and choose “Connect for Researchers.” Login with the same email and password you used previously.

When you arrive at the dashboard, click the blue “Create Project” button (Figure 3.16). You will be ready to gather data in five simple steps.

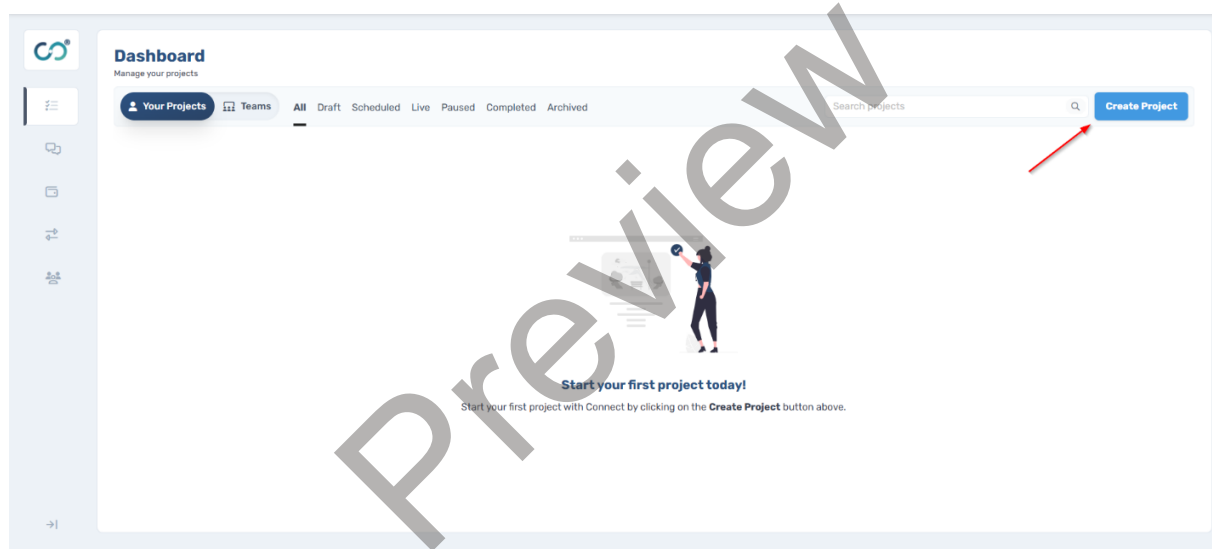


Figure 3.16. Click “Create Project” to start working on a new study.

Step 1: Basic Information

The first thing to do is provide some basic information such as a title and a description of what participants will be asked to do (Figure 3.17). Give your project the title “Social Events Survey” and describe it as “A survey about a social situation.”

Within the basic information section, you will also see checkboxes to limit the devices participants can use to take the study or to communicate whether the study requires audio, a

camera, or downloading software to participate. You can leave these settings alone for this project. Then, you are on to step two.

Figure 3.17. Basic information includes a title and project description.

Step 2: Project Link

In Step 2, you provide the link to your project (Figure 3.18). Any third-party URL will work. In this case, paste the “anonymous link” from your Qualtrics survey (see Figure 3.12 if you need a reminder of where to find this link).

Figure 3.18. Paste the link to your project within the Project Link box.

Step 3: Cost

In Step 3, you decide the details that determine how much your project will cost: how many participants you want to sample and how much you will pay them. You also need to estimate how long it will take each person to complete the project.

For this example, the recommended settings are 100 participants, with each paid 50 cents for a 3-minute study (Figure 3.19).

Figure 3.19. How many participants you want and how much you want to pay them largely determines the study cost.

Step 4: Participant Targeting

In Step 4, you specify any demographic criteria participants must meet. Connect has hundreds of data points about each participant. You can recruit people based on these characteristics.

Because this study seeks to describe people’s reactions to a moral dilemma, you will want to add a Census Matched template. To do that, select the demographic targeting option. When the popup appears, choose that you want participants from the United States and then click “Apply Census Matched Template” (Figure 3.20). The system will ask you to confirm your choice, and when you do, the quotas will automatically appear in proportion to the U.S. population. Click next and apply to confirm your selections. You are done with Step 4.

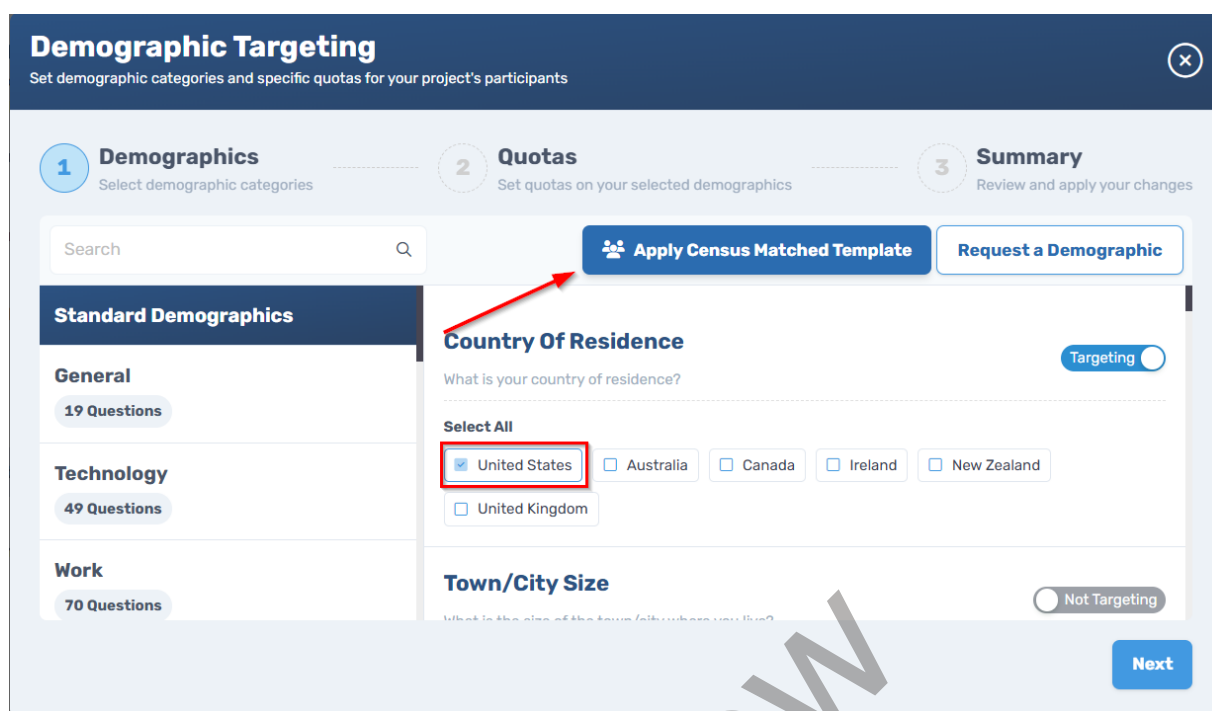


Figure 3.20. Applying the Census Matched Template.

Step 5: Completion

The last step in creating a project is determining how participants will end it. The easiest option is to use a redirect URL. A redirect URL sends anyone who completes your Qualtrics survey back to Connect. These participants get marked as “complete” and stay in a pending status for 14 days until you approve or reject their submission.

To configure a redirect URL, you need to copy the URL link provided in Connect (Figure 3.21). Then, you need to paste this link into Qualtrics. Navigate to the bottom of your Qualtrics survey and click on the “End of survey” block. A side panel will open to the left where you can change the “End of survey message” from “Default” to “Redirect to URL” (Figure 3.22). Finally, paste the URL from Connect into the website URL box. Your study is all set.

Select “Save changes.” The project will appear as a draft on your dashboard where you can launch it if your idea is chosen in class.

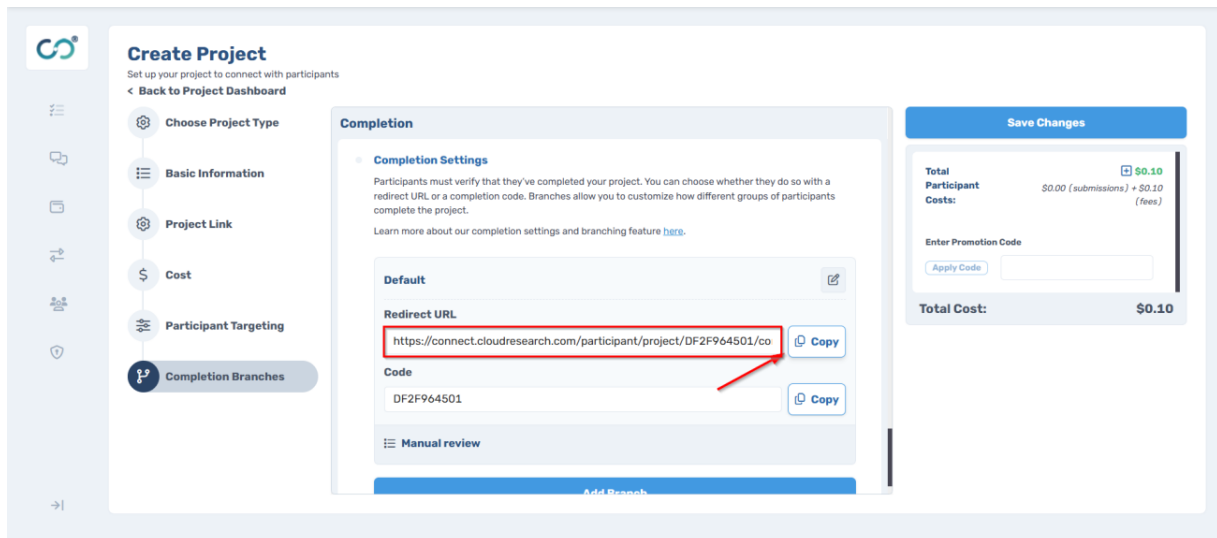


Figure 3.21. Connect provides researchers with a redirect URL to be used on the last page of the study. This URL directs participants who complete the study in Qualtrics back to Connect.

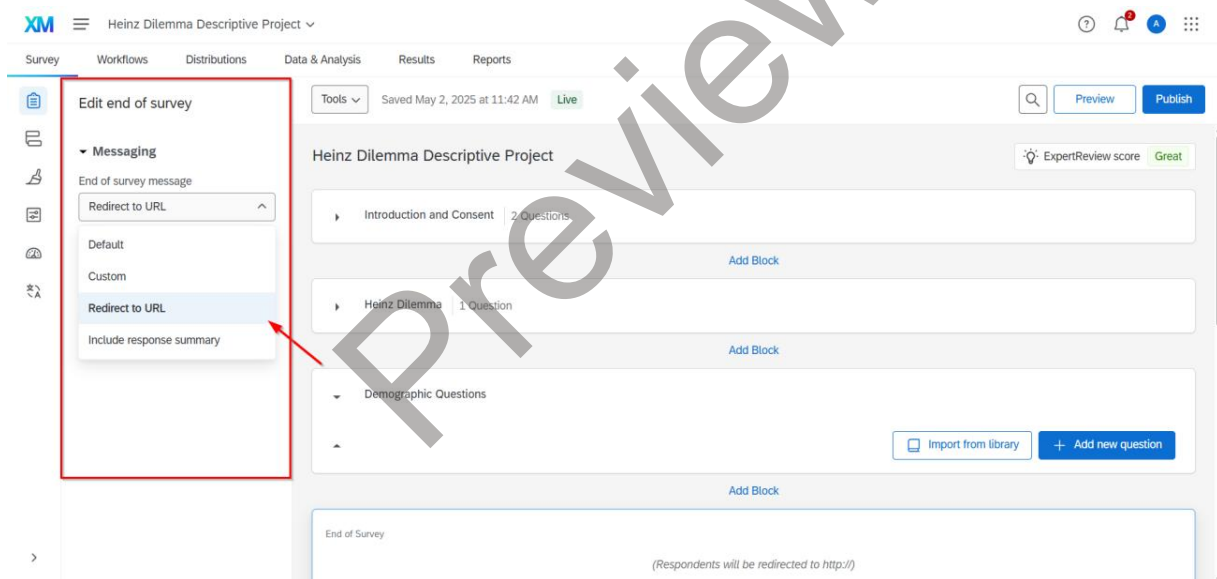


Figure 3.22. Choosing the “Redirect to URL” option within Qualtrics provides a place to paste the redirect link from Connect.

Summary

Throughout this chapter, you have explored the essential components of descriptive research. You began by examining **measurement**, learning how behavioral scientists transform abstract concepts like anxiety or moral reasoning into quantifiable variables through operational definitions. Then, you explored various **methods of descriptive research**, from large-scale

surveys like the National Survey on Drug Use and Health to observational studies using environmental sensors. Each approach provides a different window into human behavior.

Next, you gained practical experience by working with the **Heinz dilemma**, learning to program a survey in Qualtrics, analyze data in SPSS, and interpret findings about moral reasoning. The discrepancy between people's binary choices and their ratings of moral acceptability reveals the complexity of human judgment—a complexity that can only be uncovered through careful descriptive research.

Finally, you created your own descriptive study by modifying the Heinz dilemma, generating hypotheses, and preparing to collect data from a real source of research participants. This process mirrors what professional researchers do every day.

Remember that while description is important it is rarely the end goal. Like the Babylonian astronomical records that eventually led to sophisticated theories about the cosmos, today's descriptive research lays the groundwork for tomorrow's explanatory theories. The patterns scientists observe in data often raise new questions: Why do these patterns exist? What causes them? How might they be changed?

As you continue your journey through this book, you will build upon the foundation of descriptive research to explore correlational and experimental methods that help answer these deeper questions. But don't forget that it all begins with careful description—the simple yet profound act of systematically observing and recording what people think, feel, and do.

Chapter 4: Measurement

Imagine scrolling through social media and seeing this headline: "Nearly Half of Black Americans Don't Think It's Okay to Be White." The story cites a major polling company. The statistics seem clear. The methodology appears sound. Would you question it? Or would you, like millions of others, accept the headline as fact?

This isn't a hypothetical scenario. In February 2023, Rasmussen Reports released a poll showing that only 53% of Black Americans agreed with the statement, "It's okay to be White" (Rasmussen Reports, 2023). The poll spread rapidly through social media and news outlets, fueling debates about racial divisions in America. Those debates intensified when the creator of Dilbert—one of America's most popular comic strips—used the poll to make incendiary statements that eventually ended his career (Medina, 2023). But there was a fundamental problem: the survey question didn't measure what it claimed to measure.

When this textbook's authors conducted a follow-up study, we discovered that people who disagreed with the statement weren't expressing negative attitudes toward White people. Instead, they were confused by the question's meaning or responding to its use as a political slogan. When we redesigned the question to be more clear, the supposed racial divisions disappeared. As one respondent wrote, echoing the sentiments of most participants in the survey, "Color should not matter in this day and age we are all the same inside" (Hartman et al., 2023).

Although this story is about polling and race relations, it is also a reminder of why measurement matters. In behavioral research, the questions researchers ask and how they ask them have consequences. Poor measurement doesn't just produce bad data—it can shape public opinion, mislead policy makers, and affect how people think about others. As you learn about measurement in this chapter, keep this case in mind. It illustrates three important principles:

1. Single questions rarely capture complex attitudes accurately.
2. How you ask a question matters as much as what you ask.
3. Good measures need to be systematically tested for reliability and validity.

While these principles might seem like technical details, this chapter's opening shows why they matter. Getting measurement right is about more than good science—it's about responsibility to truth and the people influenced by a study's findings.

The modules in this chapter will take you through the process of finding, creating, and validating measures for behavioral research. In Module 4.1, we will show you how measurement scales work and how researchers transform abstract psychological constructs into variables they can analyze. Module 4.2 guides you through finding existing measures and creating your own, including how to use AI tools to generate and refine measurement items. In Module 4.3, you will learn how researchers test their measures for reliability and validity. Finally, Module 4.4 explores the different types of measurement scales (nominal, ordinal, interval, and ratio) and how these measurement scales affect data analysis.

By the end of this chapter, you will know how to both find and create measurement instruments for behavioral research. This skill is valuable not just for academic research but for anyone who needs to gather systematic information about how people think, feel, and behave. So, let's take a measured step in your research journey!

Chapter 4 Outline: Measurement

Module 4.1: Measurement in the Behavioral Sciences



Laboratory

Work with psychological instruments to understand how researchers measure constructs like anxiety, depression, and trauma, and what these scores reveal.

- Learn about measurement scales in behavioral research
- Work with data and compute scale scores

Module 4.2: Finding and Creating Measurement Scales



Laboratory

Discover how to locate existing measurement tools, evaluate them, and use AI to create your own scales with well-crafted items.

- Learn how to find existing measurement tools
- Use AI to create your own measurement tools

Module 4.3: Reliability and Validity



Reading

Explore the concepts of reliability and validity using real-world examples like the GAD-7 scale

- Read about the concepts of reliability and validity
- Learn how researchers validate their instruments

Module 4.4: Scales of Measurement



Laboratory

Learn about the four scales of measurement—nominal, ordinal, interval, and ratio—and why they matter

- Explore different scales of measurement
- Change scales of measurement in a dataset

Measurement in the Behavioral Sciences

Module 4.1

Work with psychological instruments to understand how researchers measure constructs like anxiety, depression, and trauma—and what these scores reveal.

In Chapter 2, we introduced tools for measuring four kinds of data in the behavioral sciences: opinions and attitudes, cognitive performance, physiological responses, and behavioral tracking (see Figure 2.10). While each kind of measurement is valuable, most studies in the behavioral sciences focus on measuring people's opinions and attitudes. Therefore, in this chapter, we focus on measurement instruments that take the form of a questionnaire, called **scale instruments**.

A **scale instrument** is a collection of carefully designed questions that work together to measure a psychological construct. Creating these instruments is both an art and a science. It is an art because researchers must find creative ways to capture complex human experiences through carefully constructed questions; it is a science because these measures must be systematically tested to ensure they are both reliable and valid.

We will begin our discussion of scale instruments with the Generalized Anxiety Disorder scale (GAD-7), introduced in the previous chapter. We will use the GAD-7 as an example of how behavioral scientists measure psychological constructs and how measurement instruments work in practice.

As you will recall, the GAD-7 was designed to assess anxiety. It consists of seven questions that ask how often a person has been bothered by specific problems over the past two weeks. Each item is scored on a scale from 0 ("Not at all") to 3 ("Nearly every day").

For example, one item asks: "Over the last 2 weeks, how often have you been bothered by feeling nervous, anxious, or on edge?" If someone responds, "Not at all," they receive 0 points. If they say, "Several days," they receive 1 point. A response of, "More than half the days" equals 2 points, and "Nearly every day" equals 3 points.

To calculate a person's total anxiety score, a researcher would sum the scores from all seven items. This creates a total score can range from 0 to 21, with higher scores indicating more severe anxiety. This summed score becomes the operational definition of anxiety.

Table 4.1 shows GAD-7 responses from 10 hypothetical participants. Each row represents one person, and each column represents their response to one of the GAD-7 items. The rightmost column shows total anxiety scores—the sum of people’s responses across all items.

Participant	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Total Anxiety Score
1	0	0	1	0	0	0	0	1
2	2	1	2	1	0	1	3	10
3	3	2	3	2	1	2	3	16
4	1	0	1	0	0	0	1	3
5	0	0	0	0	1	0	0	1
6	2	2	3	1	2	2	2	14
7	0	0	0	0	0	0	0	0
8	1	1	2	1	0	1	1	7
9	3	2	3	3	2	2	3	18
10	1	1	1	0	0	1	1	5

The sum of each person's scores represents the operationalization of anxiety

Table 4.1: GAD-7 Responses and Total Scores for 10 hypothetical participants.

Looking at the table, you can see considerable variation in anxiety. Participant 7 reported no anxiety symptoms (total score = 0), while Participant 9 reported very high anxiety (total score = 18). This variation allows researchers to examine patterns and relationships among people.

In the research activity below, you will learn how scale instruments are used for research. You will work with a dataset of 500 real participants who filled out the GAD-7, along with three other measurement scales. You will learn how to use SPSS to calculate the total score for each participant, how to conduct descriptive statistics on the sample, how to characterize the distribution of anxiety in the sample, and how to create a histogram describing the distribution of anxiety. Through this hands-on experience, you will see how measurement scales allow researchers to measure psychological concepts.

Research Activity 4.1: Working Measurement Instruments

For this exercise, you will analyze data from four validated instruments that measure depression (PHQ-9; Kroenke et al., 2001), anxiety (GAD-7; Spitzer et al, 2006), trauma (PC-PTSD-5; Prins

et al., 2016), and sleep disturbances (ISI; Bastien et al., 2001). Each of these measurement scales assesses a clinical outcome, with higher scores indicating a higher level of disorder. For example, people with higher scores on the depression scale have higher levels of depression, and people with higher scores on the sleep disturbance scale have worse sleep compared to those with lower sleep disturbance scores.

To get started, download the SPSS data file "RITC_DATA_CH04_Measurment.sav" from the *Research in the Cloud* OSF page: <https://osf.io/a8kev/>. Once the file is open, you will see 500 rows of data, with each row representing one respondent. You will also see columns that correspond to each question that was asked in Qualtrics. Figure 4.1 shows several scores from the GAD-7.

Each column represents a single PSQ-9 question about anxiety

Each row represents a participants' data

1: StartDate	ANX_1	ANX_2	ANX_3	ANX_4	ANX_5	ANX_6	ANX_7	Sleep1_1
20-Jan-2025 18:49:38								
1	3	3	3	3	3	3	3	1
2	1	1	1	1	1	1	1	1
3	2	2	2	2	2	2	2	4
4	1	1	1	1	0	1	1	4
5	0	0	0	0	0	0	0	1
6	1	1	1	3	0	3	1	2
7	0	0	0	0	0	0	0	1
8	0	0	0	0	0	0	0	1
9	1	0	0	1	0	1	0	2
10	0	0	0	1	1	0	0	5
11	0	0	0	0	0	0	0	1
12	0	0	0	0	0	0	0	1
13	0	0	0	0	0	0	0	2
14	0	0	0	0	0	0	0	1
15	0	1	1	0	0	0	0	2
16	0	0	0	0	0	0	0	3
17	0	0	1	0	0	0	0	2
18	0	0	0	0	0	0	0	1
19	0	0	0	0	0	0	0	1

Figure 4.1 Seven anxiety scale columns in the Clinical Study SPSS file.

Calculating Total Scores: Anxiety, Depression, Trauma, and Sleep Disturbance

To calculate the anxiety score for each respondent, you need to add all seven scores together to obtain a total anxiety score. The instructional video available online guides you through creating the total anxiety score and generating a histogram to examine the distribution: [video link]. You can also follow the HOW TO instructions in Box 4.1.

HOW TO: Calculate Anxiety Scores in SPSS

Follow the steps below to calculate total scores for the anxiety scale. Then, you can apply the same steps to calculate scores for the depression, sleep, and trauma scales.

Open the dataset

- Open SPSS and navigate to File --> Open --> Data
- Find the "RITC_DATA_CH04_Measurement.sav" file from where you downloaded it

Create the total score variable

- Click "Transform" in the top menu
- Select "Compute Variable..."
- In the "Target Variable" field, type "TotalAnxiety"
- In the "Numeric Expression" field create the following formula "SUM (Anx_1, Anx_2, Anx_3, Anx_4, Anx_5, Anx_6, Anx_7)"
- Click "Ok."

Verify the calculation

- The new variable should appear at the end of the dataset
- Examine or "spot check" a few cases to ensure the accuracy of the scores

Create a histogram to visualize the data

- Click "Graphs" in the top menu
- Select "Chart Builder"
- In the gallery at the bottom left, select "Histogram"
- Drag the histogram icon into the preview area
- Drag the "TotalAnxiety" variable to the x-axis (horizontal) box
- Click "Ok" to create the figure

Box 4.1. How to Calculate Anxiety Scores in SPSS

Your histogram for the GAD-7 should look like Figure 4.2. This histogram reveals an important pattern in the data known as a positive skew. A positively skewed distribution means that most people reported low anxiety (scores near 0), while progressively fewer people reported high levels of anxiety. In other words, the "tail" of the distribution extends to the right. This pattern suggests that although most people experienced minimal symptoms of anxiety, a smaller subset reported moderate to severe anxiety (scores of 15–21). Unlike the marathon completion times we discussed in the previous

chapter—which were approximately normally distributed—clinical outcomes such as anxiety, depression, and many other variables often follow a positively skewed distribution.

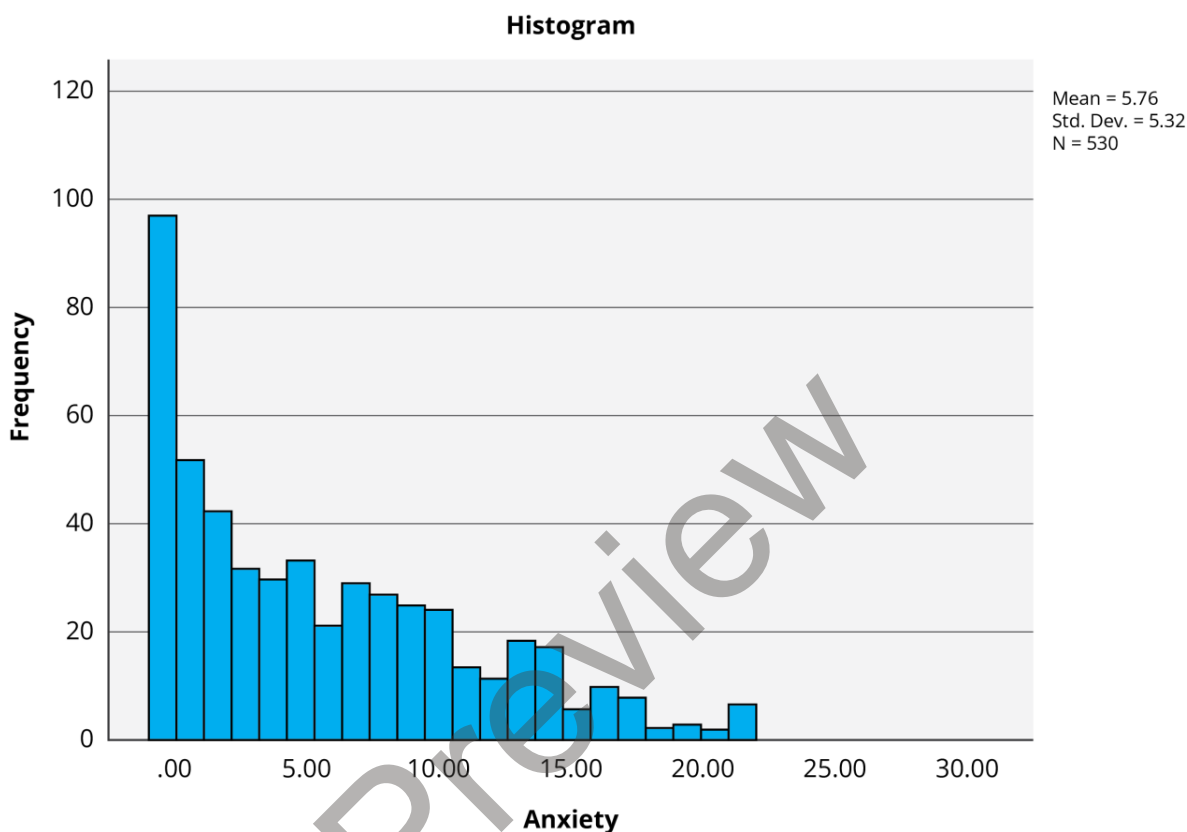


Figure 4.2. The distribution of anxiety scores, as measured by the GAD-7.

Visualizing the Distributions of Anxiety, Depression, Trauma, and Sleep Disturbance from an Online Sample of 500 People

Once you have calculated anxiety scores, repeat the process for depression, trauma, and sleep disturbances. Then, create a histogram to visualize the distribution of each variable.



Reporting on the Descriptive Statistics of Anxiety

After you have generated the histograms for the four variables in the dataset, paste them into your portfolio and write a few sentences about what the histograms show about their distributions. What does the distribution of depression look like? Is the distribution

normal or skewed? Why do you think clinical variables, such as anxiety, depression, trauma, and sleep disturbance, are positively skewed?

What the Measures Reveal

Now that you have completed the exercise, let's examine what you learned about measurement.

Through calculating total anxiety scores, you learned how researchers combine multiple items to measure psychological constructs (e.g., anxiety). Each GAD-7 item captures a different aspect of anxiety, and together they provide a more complete picture of the construct than any single question could.

This exercise also demonstrated the essential connection between theoretical constructs like anxiety and measurable data. Without measurement instruments like the GAD-7, researchers could not study psychological phenomena like anxiety, visualize their distributions, or make meaningful comparisons between people and groups. Anytime researchers want to measure a psychological construct, they need scale instruments like the GAD-7 that help transform people's experiences into a precise measurement.

Now that you understand how to measure a variable, let's discuss how to find measurement scales. There are two ways. You can use existing scales or create your own.

Finding and Creating Measurement Scales

Module 4.2

Discover how to locate existing measurement tools, evaluate them, and use AI to create your own scales with well-crafted items.

Anytime you want to measure a psychological construct, you should start by looking at what measures already exist. Behavioral scientists have developed thousands of validated measures over the years, and many are freely available. There is no need to reinvent the measuring stick.

One place to look for existing measures is in a dedicated database. Several of these exist, such as Psychology Tools (<https://psychology-tools.com>) for clinical scales and PsyToolkit (https://www.psytoolkit.org/survey-library/#google_vignette) and PsyTests (<https://psytests.org>) for general scales. These websites provide a free library with many validated measures.

The measures in these databases are often organized into categories that reflect different areas of research. For example, on PsyToolkit, under the category of personality, you can find measures of the Big Five we explored in Chapter 1, as well as questionnaires that measure characteristics like impulsivity and the need for cognition (a measure of how much people like to think). In the mental health category, there are validated measures for anxiety, depression, stress, and well-being. There are also categories for social behavior, political attitudes, consumer decision-making, and several other topics.

Beyond measurement repositories, many colleges and universities have subscriptions to databases such as PsycTests and PsycINFO. PsycTests provides access to a wide range of psychological tests and measures, many of which are available for direct use. Each entry includes a description about the purpose of the test, how it is scored, and what it measures, along with background information on its development, including reliability and validity data. PsycINFO, on the other hand, provides access to peer-reviewed publications. Like Google Scholar, PsycINFO helps you find scholarly articles, books, and dissertations that describe specific measures. By searching for keywords like “self-

esteem scale” or “anxiety questionnaire,” you can find research that discusses how these measures were developed, validated, or applied.

Research Activity 4.2: Finding Existing Measures



The purpose of this activity is to help you about measurement databases and how to find questionnaires that are of interest to you.

Start by visiting PsyToolkit (https://www.psytoolkit.org/survey-library/#google_vignette). Navigate to the library of questionnaires and find a topic you are interested in. There are hundreds of options like social media use, romantic jealousy, sleep quality, political attitudes, or how connected people feel to nature.

Once you have a topic, spend 15 minutes or so browsing the available measures. When you “run the demo” associated with a measure, you will be able to read through the items. As you explore measures, note which aspects of your topic each scale measures, and which scales seem most useful for the research you may want to do.



Reporting on the Measure You Found

After you have found a measure and read about it, paste a reference to the instrument in your research portfolio. Then, describe the topic you explored and what kind of measures you found. How did you settle on the instrument you chose? What does it measure, how many items does it contain, and what kind of scale do participants use to respond?

After describing your measure, write a few sentences explaining how you could use this instrument in a study. How would factors like scale length or response format affect your research?

Creating New Measures with AI

An effective way to measure what you are studying is to use existing measurement instruments. There are times, however, when this won't work. The measure you need might not exist, or the ones you find might not suit your research needs. For example, an

existing instrument might be too long, it might fail to capture the specific aspects of the construct you are interested in, or it might use outdated language. When you encounter these challenges, it is time to create your own measure.

Creating a measure requires decisions that shape how your scale works. Three of these decisions are: how many questions to ask, what kind of questions to ask, and how to label the response options.

How Many Questions?

One of the most important decisions when creating a measure is determining how many items to include. While it might seem simple to ask a single question, we have already seen the peril in that approach—remember the example that opened this chapter. To accurately measure a construct, the typical practice is to construct measures with multiple items. You have seen this approach with the TIPI, which uses two items per personality trait, and the GAD-7, which uses seven items to assess anxiety.

Using multiple items allows researchers to capture different aspects of the construct being measured. In the case of anxiety, one person might experience physical symptoms like restlessness while another experiences psychological symptoms like excessive worry. A scale is more likely to capture these different experiences when it uses multiple items.

Determining the optimal number of items for a scale requires a consideration of tradeoffs. If a scale has too few items, it may fail to capture important aspects of the construct. If, on the other hand, the scale has too many items it is harder for participants to complete.

For the scale you will create, we recommend five to ten items. This range typically provides enough coverage to capture the essential aspects of the construct you are interested in while remaining practical to administer and validate. As you progress in your research career, you will encounter situations that call for both shorter and longer

measures, but five to ten items is a solid starting point for scale development in this course.

What Kind of Questions?

Just as important as the number of items, creating a measurement scale requires you to choose which type of questions to use. This decision affects how well the scale will capture the construct being studied.

Table 4.2 shows the main types of questions used in measurement scales. Some are simple, like yes/no questions. Others use frequency scales that measure how often something occurs, or rating scales that capture degrees of intensity.

Among the options, Likert (pronounced “lick-ert”) items have become especially popular in behavioral science. Developed by psychologist Rensis Likert in the 1930s, these items present statements that participants rate their agreement with. What makes Likert items especially useful is that nearly any question can be transformed into a Likert item while maintaining its core meaning. This means researchers can write multiple questions that maintain a consistent response format, making the scale easier to create and easier to administer.

Question Type	Description	Example
Yes/No	Simple binary choice	"Have you ever had a flu shot?"
Multiple Choice	One answer from several options	"What is your primary mode of transportation?" [Car, Bus, Train, Bicycle, Walking]
Rating Scale	Numerical rating of intensity	"On a scale of 1-10, how satisfied are you with your job?"
Frequency Scale	How often something occurs	"How often do you exercise?" [Never, Rarely, Sometimes, Often, Very Often]
Agreement Scale (Likert)	Agreement with statements	"Exercise helps me feel better" [Strongly Disagree, Disagree, Agree, Strongly Agree]

Table 4.2. Types of questions commonly found in research studies.

When you begin constructing your scale, we recommend starting with Likert items. They

balance flexibility and simplicity, and they work well for measuring attitudes, beliefs, and experiences. As you draft items, try writing each one as a clear statement that participants can agree or disagree with.

Which Response Options?

After deciding on the kind of questions to use, you must decide how participants will respond to them. This involves both the number of response options and how they are labeled.

Typically, each response option is assigned a number. You have encountered a few different response formats in this course. In the TIPI, for instance, the scale looked like this:

Strongly Disagree (1)	Disagree (2)	Somewhat Disagree (3)	Neither Agree nor Disagree (4)	Somewhat Agree (5)	Agree (6)	Strongly Agree (7)
------------------------------	---------------------	------------------------------	---------------------------------------	---------------------------	------------------	---------------------------

In the GAD-7, by contrast, participants were given just four response options:

Not at all (0)	Several days (1)	More than half the days (2)	Nearly every day (3)
-----------------------	-------------------------	------------------------------------	-----------------------------

These numerical assignments allow researchers to calculate scores by adding or averaging responses across items. For instance, on the GAD-7, someone who responds, "More than half the days" to four items and "Several days" to three items would receive a total score of 11.

When choosing response options for your scale, consider two factors. First, how precise does the measure need to be? More response options allow for finer distinctions but may overwhelm respondents. Second, do you want a neutral middle point? Five- and seven-point scales include a neutral option, while even-numbered scales force respondents to lean one way or the other.

For the Likert scale you create, we recommend using a five-point response scale. This format provides enough precision for most purposes while remaining straightforward for

both administration and analysis.

Strategies for Writing Strong Scale Items

Beyond the basic decisions about length and response options, creating effective scale items requires attention to the details.

First, each item should measure one thing. Double-barreled statements combine multiple concepts and they make a mess of measurement. For instance, "I study regularly and get good grades," asks students about both study habits and academic performance. What if someone studies regularly but gets poor grades? How are they supposed to answer? Replace double-barreled statements with separate items: "I study regularly" and "I typically get good grades."

Second, use straightforward language to avoid double negatives. Instead of writing "I am not uncomfortable speaking in public," opt for clear statements like "I feel comfortable speaking in public" or "Public speaking makes me nervous."

Third, consider including reverse-scored items—statements written in the opposite direction of what you are measuring. Reverse-scored items can identify participants who are not reading carefully. For example, the TIPI measures extraversion with both "Extraverted, enthusiastic" and "Reserved, quiet." When calculating scores, responses to reverse-scored items are flipped before being combined with other items.

Chapter 13 describes more advanced issues about questionnaire design within the context of implementing online surveys. But the information above will get you started creating your own measures, which we show you how to do with AI.

Research Activity 4.3: Designing Your Own Measure

You are ready to create your own measurement scale. To streamline the effort, you will use AI as your research assistant.

The first, and perhaps hardest step is deciding what to measure. Don't worry about whether a scale for your chosen topic already exists. Just choose a psychological or behavioral characteristic that you are interested in and follow the steps below. You should read our example and tips for using AI before working on your measure.

Remember, the goal of this activity is for you to create a scale with about seven Likert items on a topic that interests you. In future chapters, you will have the option to validate the measure or use it in independent studies. You should be able to create a scale in 20 to 30 minutes.

Using AI to Generate Scale Items

Creating good items would traditionally take several weeks, if not months, doing things like brainstorming, consulting with colleagues, reading research, gathering data, and refining the questions through multiple drafts. While this approach will always be valuable, artificial intelligence tools like ChatGPT and Claude have opened new avenues for scale development. AI can generate initial items, explore different ways of asking about the construct, and identify elements researchers might have overlooked.

The key is to think of AI as a collaborator in scale development. Just as you might bounce ideas off a friend or an expert, you can use AI to generate items and critically evaluate them. AI does not replace your role in the process. But it provides a starting point you can refine based on your understanding of the construct and your knowledge of scale development. Let's look at an example.

Imagine you want to create a scale that measures academic stress. You could spend hours thinking of how stress manifests in academic settings or you could ask AI to generate some items and then use your judgment to evaluate which ones capture the construct well, which need revision, and what the AI might have missed.

Much of your success or failure hinges on knowing how to effectively "prompt" the AI. In other words, how to ask for what you want in a way that generates useful responses. Vague requests like, "Give me some items for a measure of academic stress" will

produce disappointing results. But, a detailed prompt that specifies what you are measuring and how you want to measure it can generate surprisingly useful suggestions.

Here is an example: "I am creating a scale to measure academic stress among college students. Generate 10 potential Likert items that capture different aspects of academic stress. Each item should be clear, specific, and follow best practices for scale development (e.g. no double-barreled questions, no double negatives). The items will use a 5-point response scale from Strongly Disagree to Strongly Agree."

When we put this prompt into Claude, it produced the items in Table 4.3. Take a minute to review the items. Jot down your thoughts. Are there elements of academic stress you think are missing? Are the questions flawed? Is there overlap that could be eliminated? After you think about the items, we will show you how to evaluate them with AI.

Preview

Item	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
The amount of coursework I have feels overwhelming					
I struggle to complete all my assignments on time					
I worry that my grades aren't good enough					
The pressure to perform well in school interferes with my sleep					
I struggle to understand the course material					
Thinking about my future academic goals makes me anxious					
I find myself constantly rushing to meet deadlines					
My schoolwork prevents me from getting enough rest					
My academic workload leaves me feeling exhausted					
I feel stressed about managing my academic responsibilities					

Table 4.3. Items to measure academic stress, produced by Claude in response to an effective prompt.

Evaluating Scale Items with AI

Once you have generated an initial set of items, you can use AI to evaluate them. For example, here is an effective prompt we gave Claude to evaluate the items in our academic stress scale:

"You are an expert in scale development. Please evaluate the items on this academic stress scale, looking for potential problems like double-barreled questions, unclear wording, redundancy, or any other problems that an expert would look for. For each item that needs improvement, explain the issue and suggest a revision."

In response, the AI pointed out several potential problems with the items and suggested improvements. For instance, it pointed out that the fourth item—The pressure to perform

well in school interferes with my sleep—and the eighth item—My schoolwork prevents me from getting enough rest—were redundant. It suggested removing one of the items. The AI also suggested that item ten—I feel stressed about managing my academic responsibilities—was too general and overlapped with other items. If you want to see how AI might analyze this scale, you can enter the prompt above into ChatGPT or Claude along with the ten-item scale and examine the full output. Later, you can repeat this step with your items.

Assembling Your Scale

It is your turn to create a measurement scale. Using the topic you chose, follow the process we demonstrated with the academic stress scale. Start by asking AI to generate about 10 items, being as specific in your prompt as we were in the example. Remember to specify that you want Likert items and the response scale the items should use.

Once you have an initial pool of items, use AI to evaluate them. Look at the AI's feedback but remember to trust your judgment too—you may notice issues the AI missed or disagree with some of its suggestions. Don't be afraid to work through multiple iterations, to change our prompts, and to explore your own approach.

Your final scale should include approximately seven items (plus or minus one or two). This length provides enough coverage to measure your construct while keeping the scale manageable for participants.

In the next section, we will explore how to evaluate your newly created scale for reliability and validity—key steps to make sure a measure works as intended.



Reporting on Your Work

After creating your instrument, report on your work in your research portfolio.

First, paste the items from your instrument into your portfolio along with the answer

scale. In a few sentences, describe the construct you are measuring. Then, reflect on the process of creating the instrument. Write a few sentences describing the criteria you used to retain or reject items and how you decided which answer scale to use.

Next, describe the feedback you received when evaluating your items. Share an example of how you used this feedback to improve an item and explain your reasoning for accepting or rejecting specific AI suggestions. How did this process deepen your understanding of scale development?

Preview

Reliability and Validity

Module 4.3

Explore the concepts of reliability and validity using real-world examples like the GAD-7 scale.

After creating a scale, researchers must test whether it works as intended.

Think about the academic stress scale we developed. Before using it in research, we need to answer two fundamental questions: Does it measure academic stress consistently? And does it measure academic stress accurately? The first question concerns the **reliability** of the measure; the second question concerns **validity**. Let's use a simple analogy to make both concepts clear.

Imagine you are measuring your temperature with a thermometer. A reliable thermometer will give you the same reading as long as your temperature has not changed. If you measure yourself twice in a row, you should get the same number. That's reliability (i.e., a consistent score). But the thermometer also needs to be valid—it needs to actually measure temperature, not humidity, not air pressure, and not how long you have held the thermometer under your tongue. That's validity (i.e., an accuracy score). Psychological measures must be both reliable and valid.

Figure 4.3 illustrates these concepts. If you think of the bullseye as the true value of the characteristic being measured, then the location of the dots represents different combinations of reliability and validity. When the dots are clustered together but far from the bullseye, as in the first target, the observations are reliable but not valid. A measure like this yields similar scores each time it is used, but it does not accurately assess the construct.

In the second target, the observations are neither reliable nor valid. When a measure is inconsistent, it sometimes captures a construct accurately but at other times it is off. Therefore, when a measure lacks reliability it cannot be valid. Reliability, in other words, is a precursor to validity—you cannot have a measure that is valid but not reliable.

Finally, in the third target, the observations are both reliable and valid. A measure with both qualities consistently measures the construct it is supposed to measure.

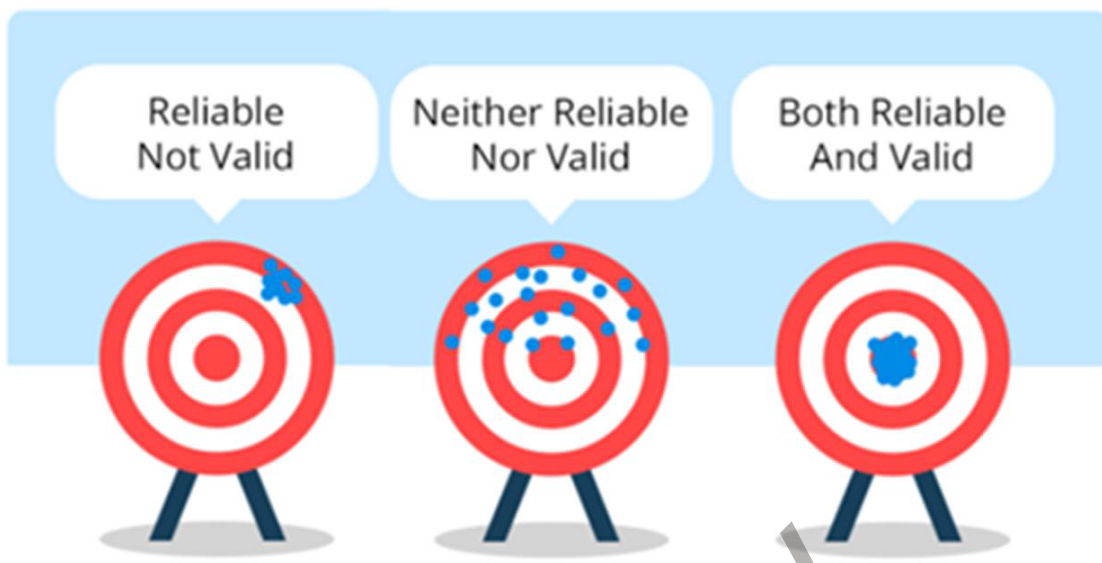


Figure 4.3. An illustration showing the relationship between reliability and validity in measurements.

An Example: Establishing Reliability and Validity of the GAD-7 Anxiety Scale

Let's look at a real example of how researchers establish reliability and validity.

When developing the GAD-7, researchers needed to show it could accurately identify people with clinical levels of anxiety. So, they collected data from nearly 3,000 people at primary care clinics in the United States (e.g., Spitzer et al., 2006). During the following week, about 1,000 of these people had a telephone interview with a mental health professional. Based on these interviews, each person was diagnosed as either having clinical anxiety or not. This diagnosis created a "gold standard" of accuracy to test against the GAD-7.

The researchers then examined how well GAD-7 scores predicted people's clinical diagnoses. They found that people diagnosed with clinical anxiety typically scored 10 or higher on the GAD-7, while those without anxiety typically scored below 10. Thus, a score of 10 proved to be an optimal cutoff point that distinguished between people who

did and did not have clinical anxiety.

But the researchers did not stop with this one assessment of validity. They also examined whether GAD-7 scores correlated with other indicators of anxiety. For example, they found that higher scores on the GAD-7 predicted more disability days, when anxiety prevented people from normal activities. Higher scores also predicted more visits with doctors and mental health professionals. People scoring higher on the GAD-7 reported greater difficulty with social relationships and more problems at work. These correlations provided additional evidence of validity because the measure predicted outcomes that a good measure of anxiety should predict.

Finally, the researchers assessed the measure's reliability. They contacted a subset of participants one week after the initial study and asked them to complete the measure again. This allowed the researchers to examine how consistent people's scores were over time. Next, the researchers examined how consistently each person responded to different items in the measure. As you will learn about next, this is a measure of the scale's internal reliability, or how well all the items in the instrument assess the same construct.

Types of Reliability

When testing if a measure works consistently, researchers examine different types of reliability. All reliability analyses rely on correlation—a statistical tool that tells how strongly two things are related.

Correlation coefficients range from -1 to +1, with +1 indicating a perfect positive relationship (as one thing increases, the other also increases), 0 indicating no relationship, and -1 indicating a perfect negative relationship (as one thing increases, the other decreases). You will learn more about correlations in Chapter 5, but within the context of reliability, researchers expect positive correlations above .70.

Just as you might check a thermometer's reliability by taking multiple readings, researchers use different techniques to verify that measures work consistently. Sometimes researchers want to

know if all the items in a measure are assessing the same construct, as we just discussed with the GAD-7. In other situations, researchers want to know if two people rating the same behavior are assigning similar scores. There are many ways to check reliability, each suited to different types of measures and research situations (Figure 4.4).

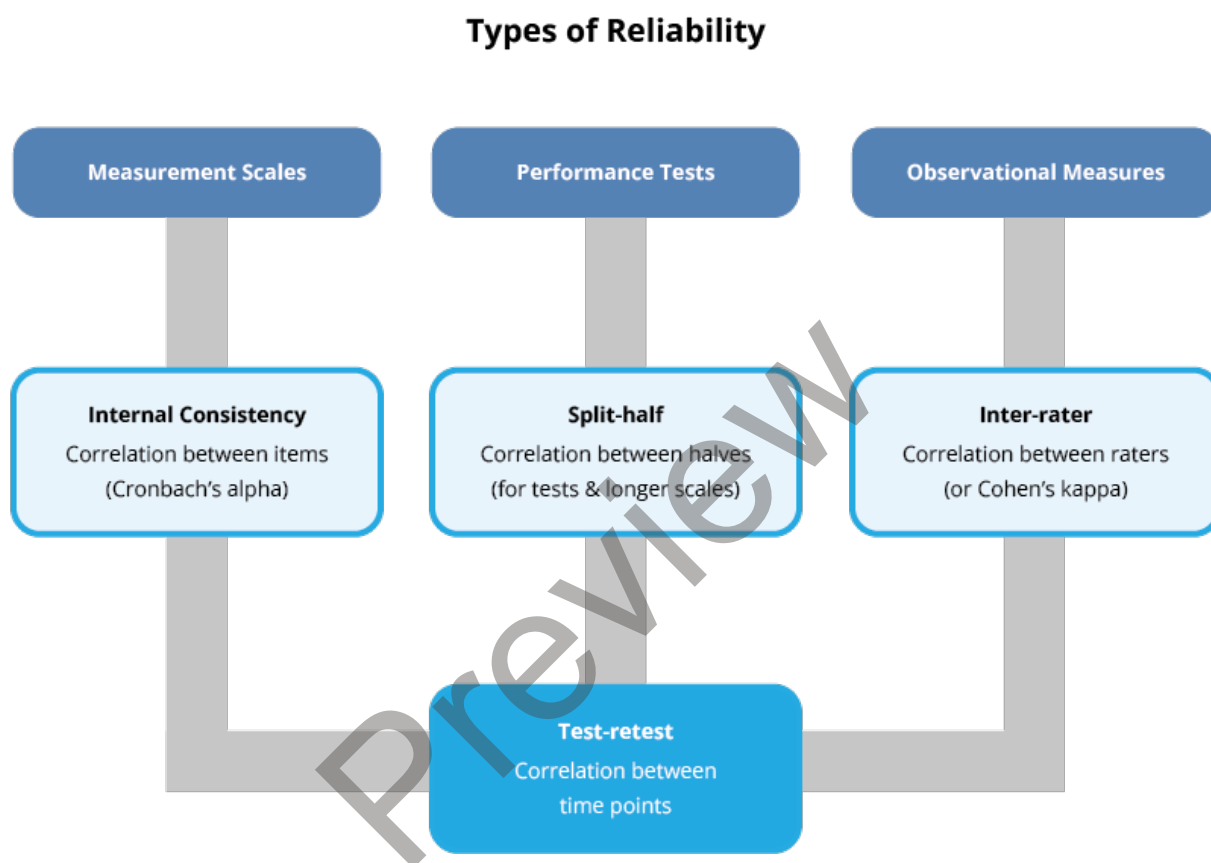


Figure 4.4 Different measures of reliability are used to evaluate different kinds of data.

Internal consistency

Let's start with the most common type of reliability: internal consistency. Internal consistency examines whether different items in the same scale correlate with each other. Theoretically, all the items should be correlated if they assess the same construct.

You can see what internal consistency assesses in Figure 4.5. The figure shows all possible correlations between the four items in a hypothetical scale (represented by the curved lines). Meanwhile, Table 4.4 depicts the strength of each correlation. For

example, Item 1 correlates with Item 2 at .65. Item 1 is correlated with Item 3 at .58, and with Item 4 at .60. In a reliable scale, the correlations should be in this ballpark or higher. If the correlations were below .30, it would suggest the items are not measuring the same thing.

Internal Consistency: All Items Correlate with Each Other

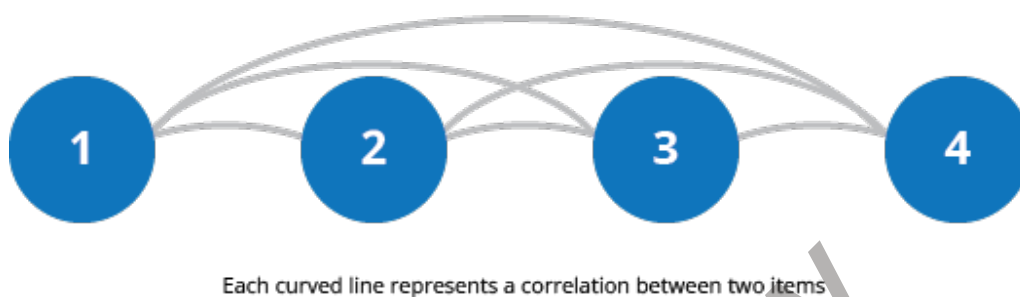


Figure 4.5. Internal consistency assesses how strongly each item within a scale is correlated with every other item.

Item	Item 1	Item 2	Item 3	Item 4
Item 1	1.00			
Item 2	.65	1.00		
Item 3	.58	.63	1.00	
Item 4	.60	.57	.61	1.00

Table 4.4. The inter-item correlations should all be moderately strong and positive in a scale where all items are measuring the same construct.

While it is useful to examine the individual correlations between items, a more meaningful metric is a single number that captures how well all the items work together. This measure is called Cronbach's alpha.

Cronbach's alpha is a statistic that ranges from 0 to 1, with higher values indicating better internal consistency. Most researchers use the conventions in Table 4.5 to interpret Cronbach's alpha.

Cronbach's Alpha	Interpretation
Above .90	Excellent
.80 to .90	Good
.70 to .80	Acceptable
Below .70	Needs improvement

Table 4.5. Cronbach's alpha scores above .70 indicate generally acceptable internal consistency (see George & Mallery, 2003).

In our example, the alpha is equal to .86, indicating good internal consistency. Thus, while the individual correlations between items were moderate (averaging .61), together the items created a reliable scale where we can be reasonably confident that each item measures the same underlying construct.

Split-half Reliability

When a scale or test is longer than about 30 items, researchers often examine split-half reliability rather than internal consistency. Split-half reliability involves dividing a measure into halves and checking how well both halves correlate.

Split-half reliability is especially common among some long personality inventories. The Minnesota Multiphasic Personality Inventory (MMPI), for example, contains over 500 true/false items. Comparing people's scores between halves helps confirm that personality is being measured and that things like the test's length or fatigue among participants is not undermining the measure.

Inter-rater Reliability

Some measures in the behavioral sciences require trained observers to rate people's behavior rather than self-reports. For instance, researchers might assess children's social skills by having observers rate playground interactions. Or, they might evaluate leadership qualities by having experts score recorded presentations. In these situations, inter-rater reliability indicates whether different observers assign similar ratings to the same behaviors.

To verify inter-rater reliability, researchers calculate a correlation between the observers' ratings. Strong positive correlations (typically above .70) suggest that the observers are applying the rating criteria consistently.

Test-retest Reliability

While each form of reliability discussed so far applies to a specific situation, researchers are interested in test-retest reliability with every measure they use. That is because test-retest reliability examines whether a measure gives consistent results over time.

To assess test-retest reliability, researchers give people the same measures on two different occasions. The time between measurements should be long enough that people do not simply remember and repeat their previous answers, but short enough that the underlying characteristic has not changed. For example, if we gave the academic stress scale to students twice during the same semester (avoiding exam periods when stress levels might change), we would expect scores to correlate. A correlation of .80 or higher between the time points suggests good test-retest reliability.

Types of Validity

While reliability assesses if a scale measures something consistently, validity assesses if it measures what it is intended to measure. At first glance, validity might seem difficult to establish. How can researchers *really* know if questions about feeling nervous or worried measure anxiety? Or if questions about feeling overwhelmed *really* capture

academic stress? The answer, as you saw with the GAD-7, lies in prediction.

If researchers have a valid measure of anxiety, then people who score high on the scale should show other signs of anxiety. In the study validating the GAD-7 this included a mental health diagnosis, visits to the doctor, and days when anxiety kept people from completing normal activities. Similarly, a valid measure of academic stress should predict relevant experiences like lower grades, more visits to academic support services, or higher dropout.

Just as researchers use correlations to establish reliability, they use correlations to test for validity. For instance, a valid anxiety measure should correlate with other indicators of anxiety. The stronger the correlations, the more confident the researcher can be that the scale measures what it is supposed to measure. There are two main approaches to validity: construct validity and criterion validity.

Construct validity

Construct validity examines whether a scale correlates with other measures in expected and theoretically meaningful ways. This involves two kinds of predictions. First, a valid scale should correlate strongly with other established measures of the same construct. For example, the GAD-7 correlates strongly with other well-validated measures of anxiety, demonstrating what behavioral scientists call **convergent validity**—the scale converges, or aligns, with similar measures of the same underlying construct.

The second kind of prediction is that a scale should not correlate (or only be weakly correlated) with measures of unrelated constructs. This is called **discriminant validity**—the scale discriminates between what it is supposed to measure and other characteristics. For example, a valid measure of math anxiety should show little to no correlation with characteristics like extraversion or artistic ability. Similarly, scores on a measure of depression should not strongly correlate with constructs like intelligence. These theoretically weak or absent correlations help confirm the scale is measuring its intended construct rather than unrelated constructs.

Criterion validity

While construct validity assesses theoretically meaningful relationships, criterion validity focuses on predicting real-world outcomes. When the predictions involve current outcomes, researchers call it **concurrent validity**; when the predictions involve future outcomes, they call it **predictive validity**.

When developing the GAD-7, researchers examined whether people's scores predicted a clinical diagnosis of anxiety. They also looked at whether GAD-7 scores predicted days missed at work or school and healthcare visits (e.g., Spitzer et al., 2006).

A strong measure has both criterion and construct validity. It correlates with related measures in meaningful ways (construct validity) and it predicts relevant, real-world outcomes (criterion validity).

Content and Face Validity: Exceptions to Prediction

Not all types of validity rely on prediction. Behavioral scientists also assess both content validity and face validity, which depend on an examination of the scale items themselves rather than their relationships with other measures.

Content validity asks whether a scale covers all important aspects of the construct it is supposed to measure. For example, there are several elements to anxiety. Some people experience anxiety as physical symptoms while others experience it more as subjective feelings. A scale that only measures one of these elements would have bad content validity. The GAD-7, on the other hand, has good content validity because its items were carefully chosen to capture different manifestations of anxiety—from psychological symptoms like worry to physical symptoms like restlessness.

Face validity simply refers to whether the items appear to measure what they claim to measure. While this might seem trivial from a scientific standpoint, it has practical implications. Participants might not take the scale seriously if it does not look like a

serious measurement tool. Face validity is also important when consulting experts. For example, an expert in the treatment of anxiety might immediately see that a scale is missing key elements in which case it would have low face validity.

Preview

Scales of Measurement

Module 4.4

Learn about the four scales of measurement—nominal, ordinal, interval, and ratio—and why they matter

There is one more thing to cover in a chapter on measurement, and it is scales of measurement.

Think about the questions you have used in research so far. When collecting demographic information, we asked about gender (male, female, non-binary, other) and ethnicity—categories without any numeric values. With the GAD-7, people rated how often they experience anxiety symptoms on a scale ranging from "Not at all" to "Nearly every day." When we ask about age, people give answers on a continuous scale (0-100). These examples show how questions exist on different measurement scales. Understanding these scales is important for analyzing the data from any behavioral study.

There are four types of measurement scales in behavioral science. You can remember them with the acronym measurement "NOIR": Nominal, Ordinal, Interval, and Ratio. Each scale has different properties that determine the comparisons and calculations you can make with the data. Table 4.5 provides an overview.

Scale	Properties	Examples from Our Projects	Other Examples
Nominal	Categories only	Gender identity (male, female, non-binary)	Major field of study
Ordinal	Order matters	TIPI responses (Strongly disagree to Strongly agree)	Course grades (A, B, C, D, F)
Interval	Equal intervals	Total GAD-7 score (0-21)	WAIS IQ scores
Ratio	True zero point	Response time in milliseconds	Age, Height

Table 4.6. Types of measurement

Nominal Measurement

The most basic type of measurement simply puts things into categories. When you select your major field of study from a list or indicate your gender identity, you are using nominal measurement. The numbers or labels assigned to these categories have no mathematical meaning—there is no sense in which "male = 1" is less than "female = 2" or that "psychology = 1" is less than "sociology = 2". All a nominal measurement says is whether things are the same or different.

Ordinal Measurement

Most questionnaire items use ordinal measurement. Any time participants rate their agreement from "Strongly Disagree" to "Strongly Agree," or rate frequency from "Never" to "Always," they are using ordinal scales. These Likert-type items are ordinal because while there's a clear ordering to the responses, researchers cannot assume the differences between responses are equal.

Consider a question from the TIPI asking people to rate their agreement with the statement "I am outgoing, sociable" using options from "Strongly Disagree" to "Strongly Agree." While there is a clear ordering to the answers—"Agree" indicates more sociability than "Disagree,"—a researcher cannot assume the differences between points are equal. The psychological distance between "Strongly Disagree" and "Disagree" might be different than the distance between "Disagree" and "Neutral," or between "Agree" and "Strongly Agree." The researcher only knows that each response represents more agreement than the last.

Interval Measurement

While individual Likert items exist on an ordinal scale, something interesting happens when they are combined into a scale score. The resulting number can be treated as an interval level measurement.

For example, remember when you summed responses across the GAD-7's seven items or computed an average across the TIPI's personality dimensions. The combined scores from these measures allowed you to compare differences between people. For example, with the GAD-7, you could say someone with a score a 15 is 5 points higher in anxiety symptoms than someone with a score of 10.

While interval measurements allow behavioral scientists to compare between people with different scores, the scale lacks a true zero point. With the GAD-7, for instance, a score of zero does not mean a complete absence of anxiety, just as an IQ score of zero does not mean a person completely lacks intelligence. The lack of a true zero point is what distinguishes interval scales from the final type of measurement: ratio scales.

Ratio Measurement

Ratio scales have a true zero point which allows researchers to know not only that Person A is 5 points higher on a scale than Person B, but also that one person's score is two or three times higher than another person's score. For example, someone who is eight feet tall is twice as tall as someone who is 4 feet tall. This makes height a ratio scale because it has a true zero point. However, IQ tests lack a true zero point—there is no such thing as an IQ of zero. Therefore, while it is possible to say that someone with an IQ of 100 has a higher score than someone with a score of 50, we cannot say the person is twice as smart.

Most measurements in the behavioral sciences exist on an interval scale, not ratio.

However, because interval and ratio scales allow the same statistical tests, researchers seldom distinguish between interval and ratio scales.

Why These Scale Types Matter in Research

Understanding measurement scales is important because it affects how you can analyze and visualize your findings in SPSS and other statistical software.

Think about the data you have seen so far. When we measured anxiety using the GAD-7, the result was numerical scores that ranged from 0 to 21. But when we measured gender identity, the result was data indicating different categories. These different measurements are handled differently when analyzing data.

Imagine you are trying to show patterns in your research findings. For interval or ratio data, like GAD-7 scores or age, histograms work beautifully. Remember the histogram we created to show the distribution of anxiety scores (see Figure 4.2)? It revealed that most people had relatively low anxiety while a few people reported high levels of anxiety. This pattern can only be seen with interval or ratio data.

But what if you tried to make a histogram of gender identity or college majors? It would not make sense because nominal categories don't have a numerical order. Instead, bar charts display nominal and ordinal data. A bar chart showing the percentage of participants in different majors, for instance, clearly displays how your sample is distributed across categories.

Even more important than visualization is choosing the right statistical techniques. When you want to understand relationships between variables—something we will explore in the next chapter—the measurement scale determines which statistical tests you can perform. For example, if you are curious about whether anxiety (interval) relates to age (ratio), you can use a correlation. But if you want to know whether gender (nominal) relates to college major (nominal), you need a different approach called chi-square analysis.

Think of measurement scales like choosing the right tool for a job. Different types of data require different analytical tools. As we move forward in this book, you will learn exactly which statistical techniques work best for different combinations of measurement scales. For now, the key is understanding that how variables are measured shapes how you can analyze them later.

Research Activity 4.4: Classify and Work with Demographic Variables in SPSS



Virtually every research study includes demographic questions. These questions help researchers understand who participated in the research and can reveal interesting patterns in the data. But different demographic characteristics are measured on different scales, which affects how the data can be analyzed. Let's explore this using the dataset from earlier in this chapter.

For this exercise, you will work with the dataset we used to examine anxiety scores, but this time we will focus on the demographic information collected from participants. Open the “RITC_DATA_CH04_Measurment.sav” data file. You will notice that along with the clinical items, the file contains several demographic variables such as age, gender, and so on. Your task is to identify the measurement scale for each demographic variable.

In SPSS, variables can be classified as Nominal, Ordinal, or Scale (which combines interval and ratio measurement). Look at each demographic question and identify which scale it exists on. For example, gender is nominal—the categories have no inherent order. Education level is ordinal—“Graduate degree” represents more education than “Bachelor's degree.” Age exists on a scale—you can meaningfully calculate averages and differences between ages. For each demographic variable, go into the SPSS variable view and label the variables according to their appropriate measurement scale. The instructional video will show you how to check and change measurement scales in SPSS or you can follow HOW TO Box 4.2.

HOW TO: Change Measurement Scales in SPSS

These steps allow you to properly classify variables by their level of measurement .

Open the dataset

- Open the "RITC_DATA_CH04_Measurement.sav" file.

Switch to Variable View

- Click on the "Variable View" tab at the bottom of the data window
- This displays all the properties of your variables in rows

Locate the Measure Column

- Scroll horizontally and find the "Measure" column
- This column displays the current measurement level of each variable

Change the Measurement Level

- Click on the cell in the "Measure" column for the variable you want to modify
- A dropdown arrow will appear when you click on the cell
- Click the dropdown arrow and choose the appropriate level of measurement
 - Choose "Nominal" for categorical variables with no inherent order (e.g., gender)
 - Choose "Ordinal" for categorical variables with a meaningful order (e.g., education level)
 - Choose "Scale" for continuous variables with equal intervals (e.g., age)

Save your Changes

- Save your dataset by clicking on "File" --> "Save"
- Your variables are now properly classified and ready for analyses

Box 4.2. How to change measurement scales in SPSS.

Summary

In this chapter, you learned about the foundations of measurement in behavioral research. We began with an example showing how a poorly designed poll question about racial attitudes led to misinterpreted headlines and the spread of misinformation. This case illustrated three core principles of measurement: complex attitudes require multiple questions, how questions are asked matters as much as what is asked, and good measures must be tested for reliability and validity.

The activities throughout the chapter reinforced these principles. First, you learned how questionnaire instruments transform abstract psychological constructs into measurable variables. Through hands-on activities with clinical scales like the GAD-7, you calculated anxiety scores and visualized their distribution, seeing firsthand how researchers assess psychological phenomena through multiple items.

Next, you explored strategies for finding existing measures in specialized databases and creating your own measures. You considered questions like how many items to include, which response format to use, and you practiced using AI to develop new measures.

The chapter emphasized two essential criteria for evaluating measurement tools: reliability (consistency) and validity (accuracy). Like a thermometer that gives consistent readings and measures temperature rather than humidity, good psychological measures must consistently assess what they claim to measure.

Finally, you examined the four scales of measurement—Nominal, Ordinal, Interval, and Ratio (NOIR)—understanding how measurement type determines appropriate analyses.

Remember that measurement is both art and science. It requires creativity to capture human experience through carefully crafted questions, and scientific rigor to make sure those questions work as intended. As you move into examining relationships between variables in the next chapter, keep in mind that the quality of your measurements directly determines the quality of your conclusions. Good science begins with good measurement.

Chapter 5: Correlational Research

Describing people's behavior is often the starting point for behavioral research. Correlational research steps in where descriptive research leaves off by aiming to uncover potential patterns of relationships between variables.

For example, in the previous chapter, we talked about measuring anxiety with the GAD-7. In a correlational study, researchers might move beyond describing anxiety to investigate how anxiety is associated with other variables such as depression, trauma, sleep, or self-esteem. Are anxious people more likely to be depressed? Are married people less likely to be anxious than unmarried people? Do people become less anxious as they get older? These are questions for correlational research.

Understanding associations, or patterns of relationships between variables, is what moves behavioral research from description to prediction. If a researcher establishes that variable A is related to variable B, then it is possible to predict values of variable B simply by knowing values of variable A. An example of this comes from the survey of mental health examined in Chapter 3. The survey, conducted by the National Institute of Mental Health, reported that younger people experienced mental illness at a higher rate than older people. This means that older people have a lower risk of developing mental illness than younger people. The ability to predict risk is a big step forward in understanding.

In this chapter, you will learn how to analyze relationships between variables and discover meaningful patterns in behavioral data. In Module 5.1, we will explore the fundamentals of correlation, including how to measure, visualize, and interpret relationships between continuous variables like anxiety and depression. You will learn to distinguish between positive and negative correlations, understand the strength of a correlation, and create correlation matrices that reveal patterns across multiple variables.

In Module 5.2, we will expand your analytical toolkit by examining different types of associations. You will discover how to analyze associations between categorical variables (like

gender or employment status) and continuous variables (like depression scores), using techniques such as *t*-tests and chi-square.

Module 5.3 offers a hands-on research project that applies correlational methods to the Heinz dilemma you encountered in Chapter 3. Using Moral Foundations Theory, you will develop and test hypotheses about how people's moral intuitions predict their ethical judgments. This guided project will walk you through each step of correlational research, from forming hypotheses to creating a study on Qualtrics, to analyzing and reporting the results.

Finally, Module 5.4 empowers you to design and conduct your own correlational study. You will develop a research question, create a survey to test the relationship between variables, and collect and analyze your own data. This independent project will allow you to apply everything you have learned so far and give you the chance to make your own discoveries about human behavior.

By the end of this chapter, you will understand how correlational research helps scientists move beyond simple description to identify meaningful patterns and make predictions about human behavior. You will also have practical experience conducting correlational studies—an essential skill for anyone interested in understanding the complex relationships that shape how people think, feel, and act.

Chapter 5 Outline: Correlational Research

Module 5.1: What Do Correlations Tell Us?



Laboratory

Learn how to measure and interpret correlations using real-world examples like anxiety and depression

- Explore what a correlation is and what it represents
- Calculate your own correlations working with real data
- Learn about conventions for the size of correlations in behavioral science

Module 5.2: Different Types of Associations



Laboratory

Explore how correlations vary across categorical and continuous variables.

- Examine what it means to test for association among different types of variables
- Conduct the statistical tests required to assess different kinds of associations
- Learn how researchers report the results of different analyses

Module 5.3: Guided Research Project: Morality and the Heinz Dilemma



Laboratory

Walk through a correlational research project examining moral decision making

- Develop hypotheses about people's moral foundations and decision making
- Analyze the data to test your hypotheses
- Interpret and report the results of the study

Module 5.4: Designing Your own Correlational Study



Laboratory

Apply your knowledge by designing, conducting, analyzing, and reporting your own correlational study

- Form your own correlational research question
- Create a survey to test the relationship between variables
- Collect, analyze, and report on your findings

What Do Correlations Tell Us?

Module 5.1

Learn how to measure and interpret correlations using real-world examples like anxiety and depression.

To understand how correlational research works, we will examine a real dataset several times throughout this chapter. The data comes from over 500 participants on Connect and is actually the same dataset you used in the previous chapter to understand the basics of measurement. Now, however, we will explore the relationships between variables rather than just examining them individually.

As a reminder, the study was created in Qualtrics and included several validated measures. You can examine the survey by downloading the “RITC_SURVEY_CH05_ClinicalStudy.qsf” file from the “Chapter 5 – Correlational Research” folder on the OSF project page:

<https://osf.io/a8kev/>. Once you have this file, upload it into Qualtrics to view the survey.

Within the survey, participants completed the Generalized Anxiety Disorder scale (GAD-7), as well as the Patient Health Questionnaire (PHQ-9), a widely-used measure of depression (Kroenke et al., 2001). Then, they completed a measure of sleep quality and experiences of trauma. In the last chapter, you calculated total scores for each of these variables. You can either use the data file you worked with previously or download a new data file from the Chapter 5 project folder, in which these total scores have been calculated for you. The file name is: RITC_DATA_CH05_ClinicalStudy.sav.

Understanding Correlations

The correlation between variables is expressed as a statistic, called a correlation coefficient. Pearson’s r is the most common correlation coefficient, and it provides information about both the direction and strength of the relationship between two variables.

Positive Relationships

Positive correlations occur when an increase in one variable predicts an increase in another variable. Lots of variables are positively correlated. People with more education tend to earn more money (e.g., Day & Neuberger, 2002). People who spend more time on social media tend

to feel more socially isolated (e.g., Primack et al., 2017). And people who have high levels of anxiety tend to also have higher levels of depression (e.g., Löwe et al., 2008; Spitzer et al., 2006). As these examples show, in positive correlations, people who score high on one measure also tend to score high on the other. This kind of relationship is often visually depicted with a scatterplot where the trend line ascends from left to right (Figure 5.1) and is mathematically represented by a positive correlation coefficient.

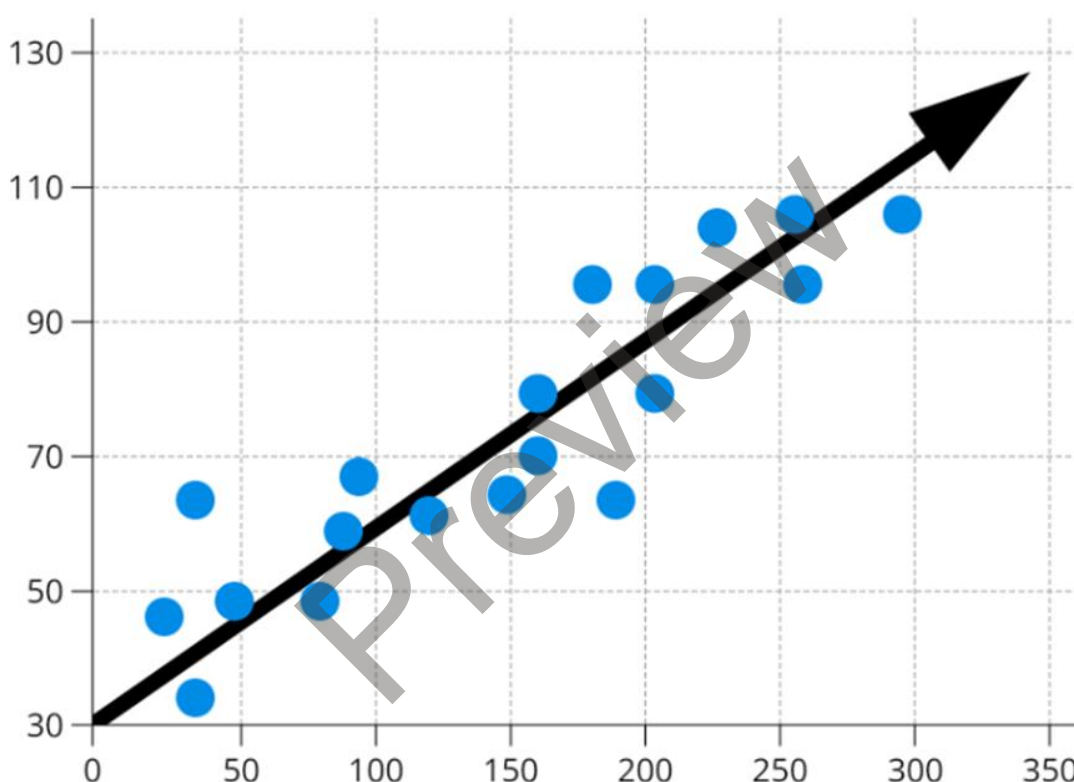


Figure 5.1. A positive correlation depicted in a scatterplot. As values on one variable increase, so do values on the second variable, yielding a line that moves up and to the right.

Research Activity 5.1: Measuring Positive Correlations



Let's examine the relationship between anxiety and depression in the clinical dataset. But first, a word of encouragement.

Many people without a strong mathematical background feel intimidated by statistics. In this case, however, the most challenging parts of the research—operationally defining variables, creating good measures, designing a study, and collecting quality data—are already complete. All you need to do is conduct some simple statistical tests with modern software that we will guide you through. There is nothing to fear.

Anxiety and Depression in the Real-World

HOW TO Box 5.1 describes how to conduct the correlation between anxiety and depression and create a scatterplot to visualize the results. You can follow the instructions in the box or watch the video we created for this project: [link to video].

The output shows a strong positive correlation ($r = .82, p < .05$). This means that as people's anxiety scores increased, their depression scores tended to increase as well. The scatterplot in Figure 5.2 shows that the points form a pattern moving upward from left to right, indicating that higher scores on one measure are associated with higher scores on the other.

HOW TO: Conduct a Correlational Analysis in SPSS

These steps will allow you to conduct a correlational analysis.

Open the dataset

- Open SPSS and navigate to File --> Open --> Data
- Find the "RITC_DATA_CH05_ClinicalStudy.sav" file from where you downloaded it

Run the Correlational Analysis

- Click on "Analyze" in the top menu
- Select "Correlate --> Bivariate"
- Move the Anxiety and Depression variables into the "Variables" box
- Make sure "Pearson" is selected under "Correlation Coefficients"
- Click "Ok" to run the analysis

Create a scatterplot to visualize the relationship

- Click on "Graphs" and select "Chart Builder"
- In the gallery, click "Scatter/Dot"
- Drag the Simple Scatter icon into the canvas area
- Drag the Anxiety variable to the x-axis
- Drag the Depression variable to the y-axis
- Click "Ok" to create the scatterplot

Add a trend line to your scatterplot

- Double-click on the scatterplot to open the Chart Editor
- Click on "Elements" in the menu bar
- Select "Fit Line at Total"
- Click "Close" to exit the Chart Editor and view your completed scatterplot

Box 5.1. How to run a correlation and create a scatterplot.

Scatter Plot of Depression by Anxiety

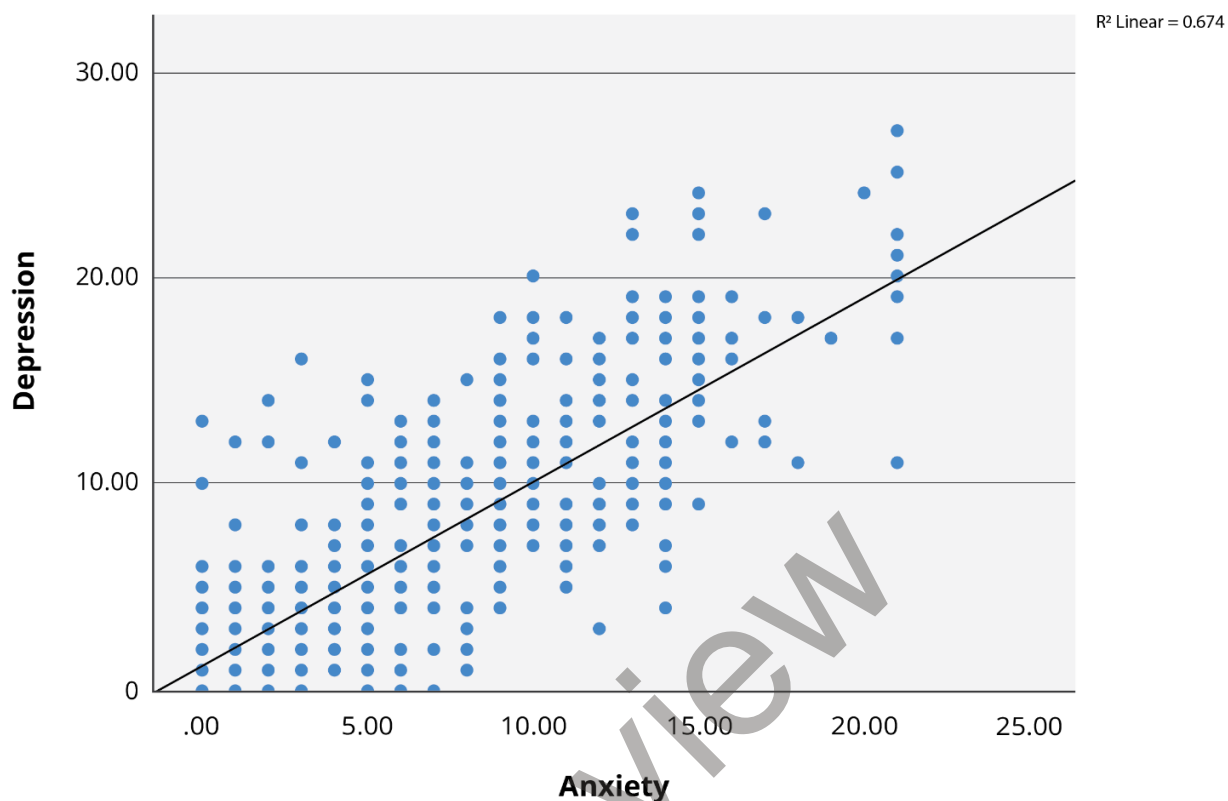


Figure 5.2. Scatterplot depicting a positive correlation between anxiety and depression. Each dot represents a single person.



Reporting the Positive Correlation between Anxiety and Depression

Once you have generated the scatterplot, paste it in your portfolio and write a few sentences about what the scatterplot shows about the relationship between anxiety and depression. Why would anxiety and depression tend to be positively correlated? In other words, what underlying factors might explain why people who experience more anxiety also tend to experience more depression?



Stop and Discuss!

Think about the positive correlational relationships you observe in your own life and in psychology. Then, discuss the following questions with classmates or friends.

1. What examples of positive correlation do you see in daily life?
2. What psychological characteristics do you think might be positively related to each other?
3. Consider how you might design a study to test whether the relationships you've identified actually exist. What would you need to measure? How would you collect the data? This kind of scientific thinking is at the heart of this chapter.

Correlations, Trends, and Prediction

When examining correlations, the predictions researchers make are not perfect for every person. You can see this by examining the scatterplot in Figure 5.2. Notice that some people have depression scores above ten but anxiety scores of zero. For these people, high depression scores do not predict high anxiety scores. Nevertheless, across the entire sample, people with high anxiety scores generally have higher depression scores and vice versa.

Clinical psychologists know that anxiety and depression usually have a strong ability to predict one another because the correlation between these variables is one of the most fundamental findings within the field (e.g., Kalin, 2020; Spitzer et al., 2006). This strong association helps explain why many mental health treatments target both anxiety and depression simultaneously.

Negative Relationships

Not all correlations are positive, some are negative. A negative correlation indicates that as scores on one variable increase scores on the other variable tend to *decrease*. This is also sometimes called an inverse relationship.

A negative correlation is how we would describe the relationship between anxiety and self-esteem. As people's anxiety increases their self-esteem generally decreases (e.g., Löwe et al., 2008). Thus, these two variables are negatively correlated.

Other examples of negative correlations abound. For example, as people's time spent multitasking increases, their satisfaction with what they are doing tends to decrease (Mark et al., 2008; Mark et al., 2016). The less time people spend sleeping, the more health problems they tend to encounter (Luyster et al., 2012). And, as people's sense of control over their lives

decreases, the frequency of depression tends to increase. When a negative correlation is graphed, the result is a line that moves down and to the right as in Figure 5.3.

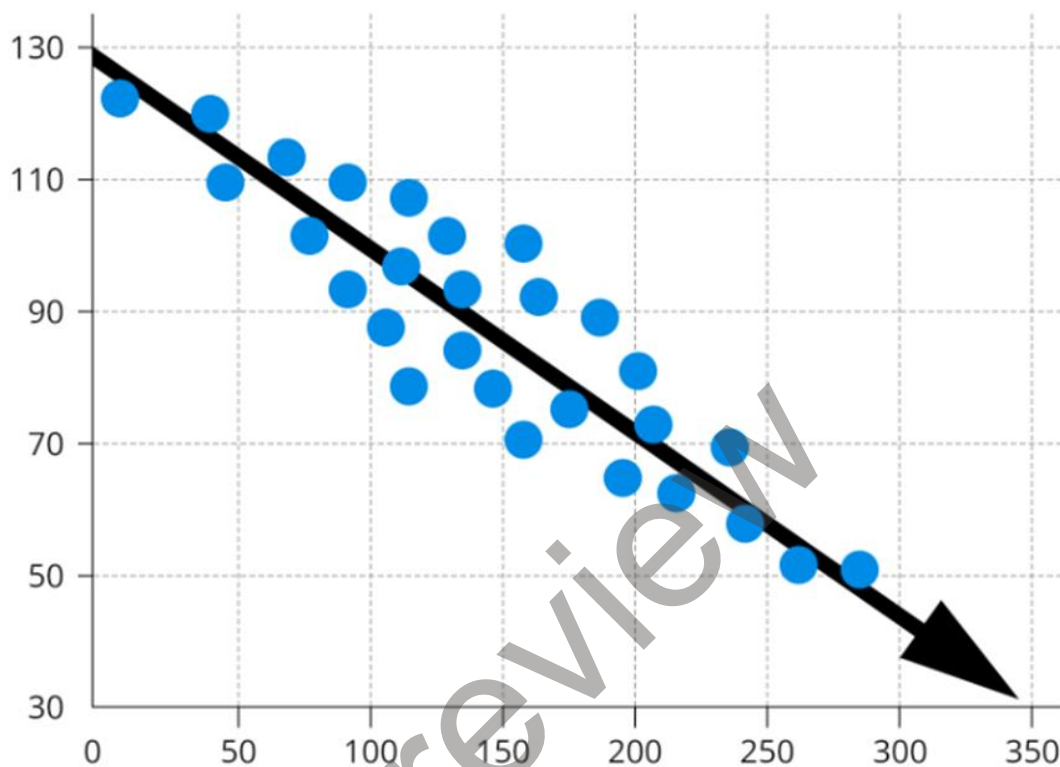


Figure 5.3. Within a scatterplot, negative correlations form a line that moves down and to the right.

Research Activity 5.2: Measuring a Negative Correlation



Let's look at a negative correlation. In Chapter 3, you saw that the National Institutes of Mental Health conducted a survey that found younger adults reported higher rates of mental illness than older adults (see Figure 3.2). Based on this finding, we might expect a similar pattern in the clinical dataset. Specifically, depression and anxiety scores might decrease with age.

To test this relationship, you can follow the same steps within SPSS that you used for positive correlation. HOW TO Box 5.1 describes how to conduct the correlation, except this time you will add age into the correlation window as well.

Age and Emotional Distress

When you conduct this correlation, you will find a negative relationship between age and depression ($r = -.20, p < .05$) and between age and anxiety ($r = -.23, p < .05$). The minus sign indicates that as age increases depression and anxiety scores tend to decrease. While neither of these relationships are as strong as the one between anxiety and depression, the negative correlations show that older participants in the sample generally reported lower levels of depression and anxiety than younger participants.



Reporting the Negative Correlations between Age, and Anxiety and Depression

Once you have conducted the analysis, paste the SPSS output to your portfolio. Write a few sentences describing what the output shows about the relationship between age and depression, and age and anxiety. Interpret the correlations in your own words. Why do anxiety and depression tend to be negatively correlated with age? In other words, what underlying factors might explain why people who are younger, such as teenagers, might experience more anxiety and depression compared to older people?

When a Correlation Counts: Magnitude and Statistical Significance

The Magnitude of a Correlation

While the sign of a correlation coefficient tells you how two variables are related, it does not say anything about the strength of the relationship. For that information, you must examine the numerical value of the Pearson's r statistic. The closer the value is to 1 or -1, the stronger the relationship. The closer the value is to zero, the weaker the relationship. And, if the correlation is close to zero there is no relationship between the variables (Figure 5.4).

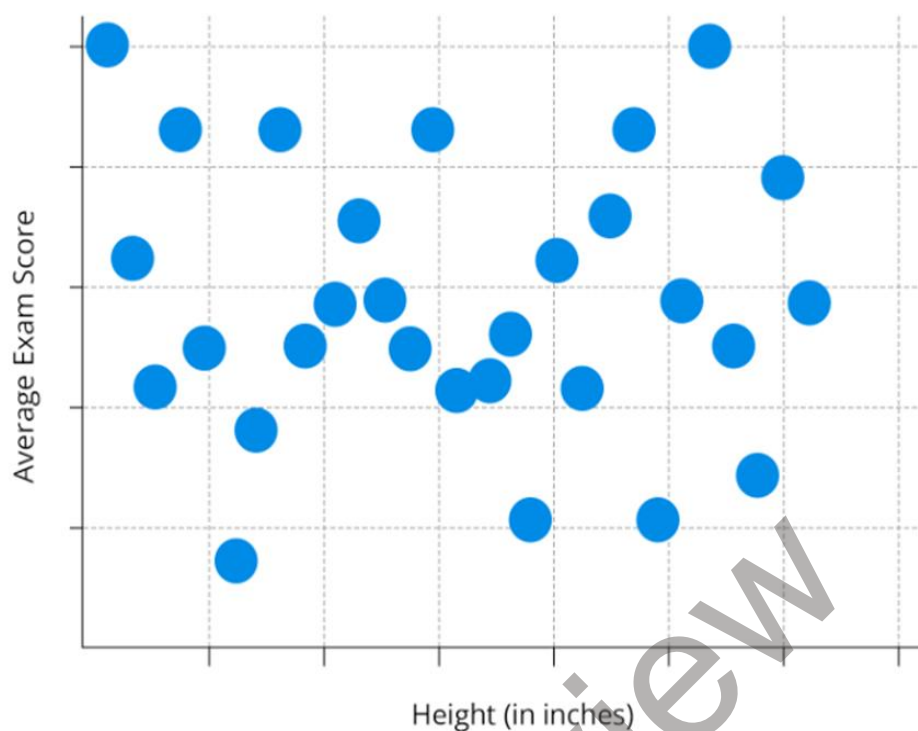


Figure 5.4. A scatterplot depicting no relationship between two variables.

Across the behavioral sciences, researchers share conventions for interpreting the size of a correlation coefficient. These conventions come from the work of Jacob Cohen, a prominent figure in the field of psychology and statistics (1992). The conventions apply to the size of a correlation regardless of its direction.

A correlation between 0.1 and 0.3 is considered small, indicating a weak association between the variables. A coefficient above 0.3 and below 0.5 is considered moderate, suggesting a more substantial but still modest relationship. Finally, a correlation of 0.5 or higher is considered large, indicating a strong relationship between the variables. From the examples you have seen, the correlation between anxiety and depression would be considered large and the correlation between depression and age would be considered small. Table 5.3 presents other examples of small, moderate, and large correlations.

r value (in absolute terms)	Effect size	Examples
-------------------------------	-------------	----------

< .30	Small	Conscientiousness and medication adherence ($r = .15$; Molloy et al., 2014)
.30 to .50	Medium	Job satisfaction and job performance ($r = .30$; Judge et al., 2001)
> .50	Large	Parental education and child's academic attainment ($r = .50-.60$; Dubow et al., 2009)

Table 5.1. Guidelines for the strength of correlation coefficients. These conventions are the same for positive and negative values.

Statistical Significance

A correlation that is close to zero indicates no relationship between variables. Yet when behavioral scientists calculate the correlation between any two variables, they rarely get exactly zero. Even if two variables should not be related—say, shoe size and empathy—the researcher will probably find some small correlation, maybe .03 or -.04.

The question then becomes: when is a correlation large enough to be meaningful rather than just random noise? The answer comes from statistical significance. Researchers use tests of statistical significance to determine if a correlation is larger than what would occur by chance. The result of these tests is expressed as a probability or " p -value".

By convention, if p is less than .05 the correlation is *statistically significant*. When $p < .05$, behavioral scientists have good reason to believe the correlation represents a real relationship rather than random variation in the data. A p -value below .05 means that if there were truly no relationship between the variables, a correlation as large as what is observed (or larger) would be found less than 5% of the time by chance alone.

In the data you examined earlier, all the correlations were statistically significant. The strong positive correlation between anxiety and depression ($r = .82, p < .05$) and the negative correlations between age and depression ($r = -.20, p < .05$) and age and anxiety ($r = -.23, p < .05$) all represent reliable patterns in the data, rather than chance findings.

The logic of statistical significance applies to all the statistical tests you will encounter in this book. Whether looking at correlations or any other statistical test, behavioral scientists use $p < .05$ as a guideline for determining whether the results are meaningful or more likely due to chance.

Research Activity 5.3: Examining a Correlation Matrix



When there are multiple variables in a study, researchers often want to examine how all the variables relate to each other. Rather than calculating correlations one at a time, it is possible to generate a correlation matrix—a table that shows all possible correlations at once.

To create a correlation matrix in SPSS, follow the steps in HOW TO Box 5.1, with one additional step. Instead of placing two variables into the “Variables” box in the correlation analysis, move all the variables of interest into the box at once. The resulting output will show every possible correlation between the variables in one table.

Figure 5.5 shows a correlation matrix with seven variables from the clinical dataset. The variables in the matrix include depression, anxiety, sleep quality, trauma, age, income, and education. Each cell in the matrix shows the correlation between the variables in that row and column. For instance, to find the correlation between depression and anxiety, look at where the depression row intersects with the anxiety column. The value there is $r = .82$.

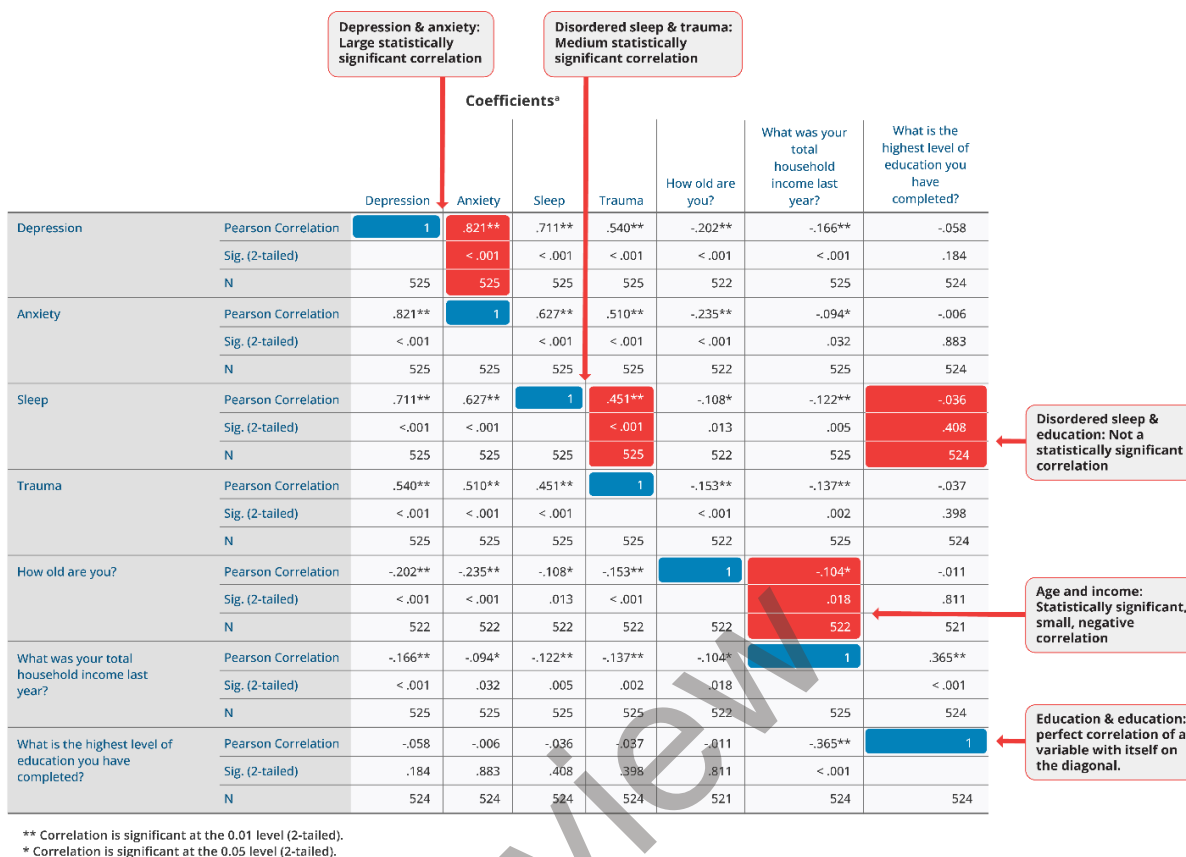


Figure 5.5 An SPSS correlation matrix showing correlations between four clinical and three demographic variables.

A correlation matrix has several key features. First, along the diagonal line running from top-left to bottom-right, you will see all 1's—these represent each variable's correlation with itself.

Second, above and below the diagonal line, the matrix is symmetrical. This means the correlation between any two variables appears twice. You can simply ignore everything below the diagonal.

Third, the "Sig. (2-tailed) value" is the *p* value for each correlation. Asterisks mark statistically significant correlations, with one asterisk indicating *p* < .05 and two asterisks indicating *p* < .01.

Finally, the "N" rows show the sample size for each correlation.

Looking at the matrix reveals patterns in how these variables correlate. For example, depression shows a strong positive correlation with both anxiety (*r* = .82) and sleep problems (*r* = .71), a moderate positive correlation with trauma (*r* = .54), and a small negative correlation with both

age ($r = -.20$) and income ($r = -.16$). The correlation between disordered sleep and education is not significant.

Reporting the Results of a Correlation Analysis

Writing about statistical results is an important part of behavioral research. It is used not only to clearly communicate the results of a study to others, but often it helps clarify a researcher's own thinking about what was found. A good write-up presents the results in clear, non-technical language that any reader can understand, while also including the essential statistical information. Let's look at an example.

To describe the correlation between anxiety and depression you might write:

"I conducted a Pearson correlation analysis to examine the relationship between anxiety and depression. I found a significant positive correlation between anxiety (as measured by the GAD-7) and depression (as measured by the PHQ-9), $r(524) = .82, p < .05$. This result suggests that anxiety and depression are strongly related. Specifically, the results suggest that people who experience higher levels of anxiety also tend to experience higher levels of depression."

In this write-up, the correlation coefficient ($r = .82$) tells readers how strong the relationship is and in what direction. The number in parentheses (524) represents the degrees of freedom, which is related to the sample size (sample size $[N]$ minus 2). The p-value ($p < .05$) indicates the relationship is statistically significant and larger than what would be expected by chance. Together, these elements give readers the information they need. Plus, the plain language interpretation helps everyone understand what the results mean in practical terms.



Stop and Discuss!

Take a minute to examine the correlation matrix in Figure 5.5. It shows the relationships among depression, anxiety, sleep problems, trauma, age, income, and education. After you've examined the correlations, discuss each of the questions below.

1. For each correlation in the matrix, practice interpreting and reporting the findings. As a class activity, go around the room with each person interpreting one of the correlations.

Start with the r coefficient and its direction, note its size using Cohen's conventions, and then explain what it means in plain language.

2. After writing your results, search Google Scholar for studies that have examined these same relationships. For instance, you can search "depression anxiety correlation meta-analysis" or "sleep depression correlation meta-analysis." How do the findings from this example study compare with what other researchers have found? Are the correlations similar in size, direction, and magnitude? What might explain any differences you find?
3. Discuss your interpretations and literature findings with your classmates. Some interesting correlations to discuss include depression and trauma, depression and age, depression and income, anxiety and sleep, anxiety and trauma, and sleep and trauma. What patterns do you notice in how these variables relate to each other? Do these relationships align with what you might have predicted?



Writing the Results of a Pearson's Correlation Analysis in a Formal Scientific Format

Once you have created the correlation matrix, paste it to your portfolio. Select two correlations, one that is significant and one that is not. Write up the results using the template for reporting correlations above. Make sure to report the r coefficient and its direction, note its size using Cohen's conventions, and then explain what it means in plain language.

Different Types of Associations

Module 5.2

Explore how correlations vary across categorical and continuous variables

So far, we have focused on correlations between continuous variables like anxiety and depression. However, behavioral scientists often want to understand relationships between variables that are not continuous. For example, is anxiety (continuous) related to gender (not continuous)? When examining the relationship between variables that are not both continuous, researchers must use different statistical tools.

Recall in the previous chapter that you learned about four scales of measurement: nominal, ordinal, interval, and ratio (remember the acronym ‘measurement NOIR’). For practical purposes, these measurement scales can be simplified into two broad categories: **categorical** and **continuous**.

We can treat both interval and ratio scales as continuous variables because they exist on a continuum and are analyzed the same way. We can also treat ordinal variables, like Likert scales, as continuous variables (even though there are specialized techniques for ordinal data that we do not cover here). Finally, we can treat nominal variables as categorical data. This yields two main types of variables—categorical and continuous—that can be associated with each other in three possible combinations (Figure 5.6).

Each type of relationship requires a different analytical approach. The type of relationship also determines how researchers visualize the results. Let's examine each type of relationship using examples from the anxiety dataset.

Measurement in Behavioral Research

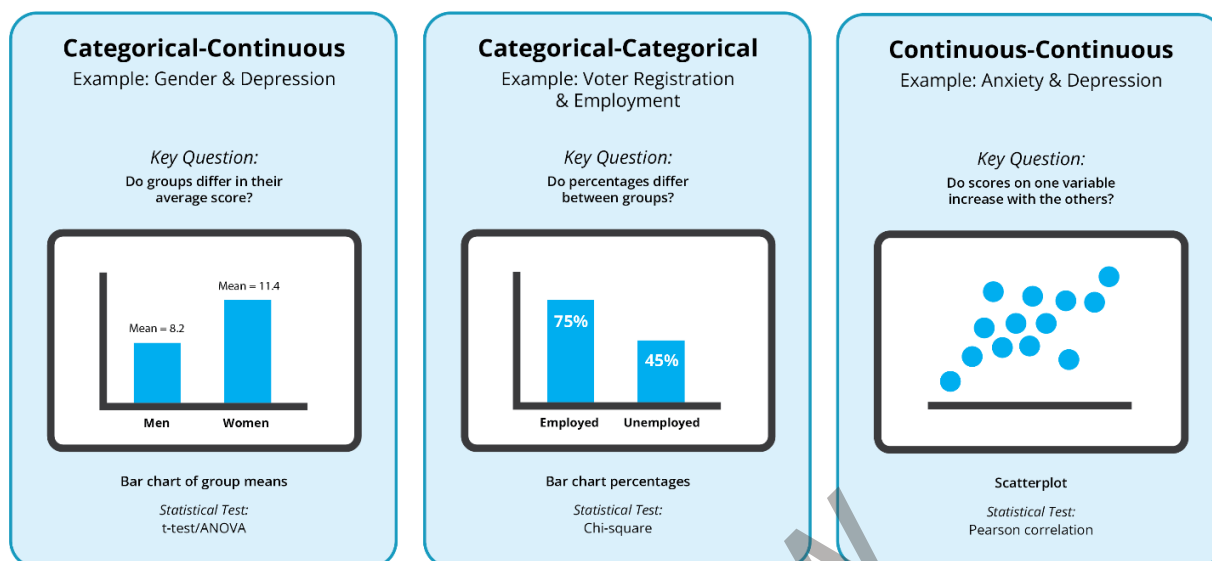


Figure 5.6. Three possible combinations of variables.

Associations between Categorical and Continuous Variables

Many research questions examine how different groups of people differ on a continuous measure. A school administrator, for instance, might want to know if students at private schools have higher test scores than students at public schools. In this case, school type (public vs. private) is categorical while test scores are continuous. Public health researchers might investigate whether people living in cities report higher stress than those in rural areas. Medical researchers might examine whether participants in exercise programs show lower blood pressure than people not in such programs. Or developmental psychologists might want to know whether first-born children score differently on personality traits than later-born children. In each of these cases, the researchers want to know if one group of people score differently on a measure than another group.

To analyze these kinds of relationships, researchers compare the average scores for each group. For example, in the anxiety dataset, we can examine whether the average depression score is higher for one gender group than another.

Research Activity 5.4: Comparing Gender Differences in Depression



In Chapter 3, we described the National Institutes of Mental Health survey that found higher rates of mental illness among women than men. This finding is an example of a relationship between a categorical variable (gender) and a continuous variable (mental health symptoms). When examining this type of relationship, researchers calculate the average score for each group and then compare the two. Let's see how this works with our clinical dataset.

Figure 5.7 shows the average depression scores for men ($M = 6.10$) and women ($M = 7.27$). The difference in average scores represents an association between gender (the categorical variable) and depression (the continuous variable). We can call this an association because knowing someone's gender gives us information to predict their depression score—as a group, women score about one point higher than men. If there were no association between gender and depression, we would expect similar levels of depression across groups.

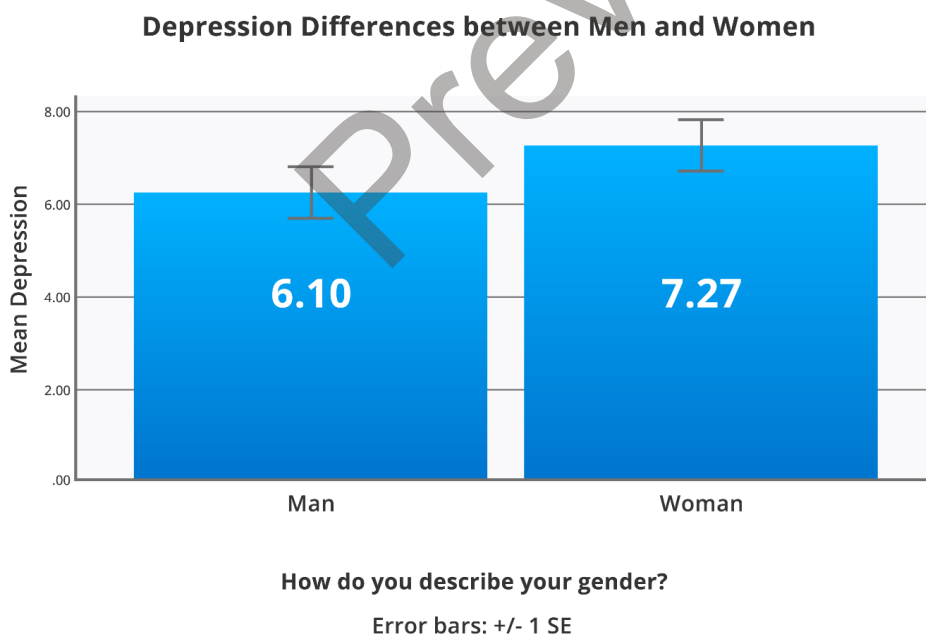


Figure 5.7. Differences in depression scores between men and women.

Conducting a t -test

Just as we used Pearson's r to describe the relationship between continuous variables, researchers use something called a t -test to determine if differences between group averages are meaningful. The t -test assesses whether the difference between two groups is larger than what would be expected by chance alone. Like with correlations, a $p < .05$ is the criterion for statistical significance.

HOW TO Box 5.2 describes how to conduct the t -test examining whether men's and women's depression scores are different on average. You can follow the instructions in the box or watch the video for this project: [link to video].

HOW TO: Conduct an Independent Samples t -test in SPSS

These steps will allow you to compare the means of two independent groups, such as the depression scores between men and women.

Open the dataset

- Open the "RITC_DATA_CH05_ClinicalStudy.sav" file

Access the t -test function

- Click "Analyze" in the top menu
- Select "Compare Means" from the dropdown menu
- Click on "Independent Samples T-Test"

Select variables for analysis

- Move the depression score into the "Test Variable(s)" box
- Move the categorical variable gender into the "Grouping Variable" box

Define the groups to compare

- After moving gender into the grouping variable box, click the "Define Groups" button
- Enter "1" into the "Group 1" box, representing men
- Enter "2" into the "Group 2" box, representing women
- Click "Continue" to return to the main dialog box

Run the analysis

- Click "OK" to execute the t -test
- Results will appear in the output viewer window
- The output includes descriptive statistics for each group and the t -test results showing whether differences between groups are statistically significant

Box 5.2. Instructions for conducting a t -test.

The t -test shows a significant difference between men's and women's depression scores ($t = 2.3$, $p < .05$), meaning the difference is larger than what would be expected by chance. This significant test confirms that there is a real, reliable relationship between gender and depression in this sample.

Just as with correlations, researchers need to communicate the findings of t -tests in a scientific report. An example of how to do this might read:

“I conducted an independent samples t -test to examine differences in depression between men ($M = 6.10$, $SD = 5.42$) and women ($M = 7.27$, $SD = 6.09$). There was a significant difference, $t(498) = 2.30$, $p < .05$, suggesting that gender is associated with depression. More specifically, my results indicated that women tend to experience higher levels of depression than men, even though the effect size (Cohen's $d = 0.20$) suggests the difference is small in magnitude.”

If you are wondering where that last bit about effect sizes came from and why $d = .20$ is considered small, continue to the next section.

Making Sense of Effect Sizes

Just as researchers have guidelines for interpreting the size of correlations, they have guidelines for interpreting the size of differences between group means. The most common statistic for this is Cohen's d , which measures the difference between two group means relative to the standard deviation of both groups.

A Cohen's d around 0.2 represents a small difference between groups. Values around 0.4 represent medium-sized differences. And values of 0.6 or larger represent large differences. These guidelines help researchers communicate not just whether group differences are statistically significant, but how meaningful they are in practical terms. Table 5.2 contains examples of research findings with different effect sizes.

In our sample project, we found a Cohen's d of 0.2, indicating a small but reliable difference between men's and women's depression scores. This aligns with the survey from Chapter 3, where women showed somewhat higher rates of mental illness than men. While the difference is small, the consistency of this pattern across different studies—from large national surveys down

to our sample of 500 participants—suggests the gender difference in depression is a reliable phenomenon, even if its effects are modest.

Cohen's d	Effect Size	Examples from Research
0.2	Small	Gender differences in depression ($d = 0.2$; Salk et al., 2017)
0.4	Medium	Social media use and anxiety ($d = 0.45$; Hunt et al., 2018)
0.6	Large	Exercise intervention effects on stress ($d = 0.84$; Anderson & Durstine, 2019)

Table 5.2. Conventions for interpreting Cohen's d effect sizes.

Overall, the key to understanding categorical-continuous relationships is recognizing that they are still just a pattern of association. However, instead of looking at how two continuous variables move together (like anxiety and depression), researchers compare the average scores of different groups. The statistical tools change—from correlation coefficients to t -tests—but the goal remains the same: examining whether scores on one variable help predict scores on another.



Reporting on Gender Differences in Depression with a t -test

Once you have conducted your t -test, paste the results into your portfolio. Report on the results using the template above.

Associations between Two Categorical Variables

Many questions in behavioral research involve relationships between categorical variables. For example, researchers might want to know whether employed people are more likely to register as voters than unemployed people. In this case, both variables—employment status (employed vs. unemployed) and voter registration (registered vs. not registered)—are categorical. Similarly, researchers might investigate whether college graduates are more likely to own homes than non-graduates, or whether people living in cities are more likely to use public transportation than those living in rural areas.

When examining relationships between categorical variables, researchers compare percentages between groups: what percentage of employed people are registered to vote vs what percentage of unemployed people are registered to vote? To examine whether percentages differ between groups, researchers use a statistic called a chi-square (χ^2). The idea behind the chi-square is the same as Pearson's r and the t -test: researchers want to know if the p -value associated with the statistic is smaller than .05. If so, there is an association between the variables.

Research Activity 5.5: Comparing Categorical Variables

Many variables are naturally categorical (voter registration: registered vs. not-registered). But researchers sometimes create a categorical variable from a continuous measure to answer a specific research question.

For example, while we previously examined the difference in average depression scores between men and women, the same data could be used to test whether women are more likely than men to experience severe depression. To do so, we would draw upon clinical cutoffs for the PHQ-9, where a score of 20 or higher indicates severe depression. Using the cutoffs, we could create two groups: one with anyone who has a PHQ-9 score of 20 or greater (the 'severe depression' group) and a second with anyone who has a score below 20 (the 'no severe depression' group). Creating these groups transforms the continuous measure into a categorical one that we can use to test gender differences in severe depression.

To create these groups and conduct the Chi-square test, you can follow the instructions in HOW TO Box 5.3 or watch the accompanying video for this exercise: [video link].

HOW TO: Compare Categorical Variables Using Chi-Square in SPSS

These steps help you analyze whether a relationship exists between two categorical variables such as gender and severe depression.

Open the dataset

- Open the "RITC_DATA_CH05_ClinicalStudy.sav" file

Create the categorical variable

- To create categories from a continuous variable like depression scores, click "Transform" in the menu bar
- Select "Recode into Different Variables..."
- Move the Depression variable into the "Input Variable" box
- in the "Output Variable" box, enter a name for the new variable such as "SevereDepression"
- Click "Old and New Values" to define your categories
 - Click "Range, LOWEST through value:" and enter "19" in the box. Then, in the "New Value" box on the right enter "1." Then click, "Add." All scores below the cutoff of 20 will now be scored as a "1"
 - Click "Range, value through HIGHEST:" and enter "20" in the box. Then, in the "New Value" box on the right enter "2." Then click, "Add." All scores above the cutoff of 20 will now be scored as a "2."
- Click "Continue" and then "OK"

Run a Chi-Square analysis

- Click on "Analyze" in the top menu
- Select "Descriptive Statistics > Crosstabs"
- Move one categorical variable, Gender, to the "Row(s)" box
- Move the other categorical variable, SevereDepression, to the "Column(s)" box
- Click "Statistics" and check "Chi-square"
- Click "Continue"
- Click "Cells" and check "Row percentages"
- Click "Continue" and then "OK"

Interpret your results

- Look at the "Chi-Square Tests" table to find the Pearson Chi-Square value and significance level
- Look at the Crosstabulation table to compare percentages between groups

Box 5.3. Instructions for transforming a continuous measure into a categorical one and conducting a Chi-square.

In our sample project, 5.5% of women experienced severe depression compared to just 0.5% of men. The chi square value was 10.7, $p < .05$, indicating that the difference in percentages was statistically significant. Across the behavioral sciences, analyses like this—where researchers transform a continuous measure into a categorical one—are common, especially in clinical settings where diagnostic categories are often defined based on numerical cutoffs.

To report the results of a chi-square analysis, you might write:

“I conducted a chi-square test to examine if there was a difference in the percentage of men and women who scored above a 20 on the PHQ-9. There was a significant difference, $\chi^2 (2) = 10.76$, $p < .05$, suggesting that gender is associated with severe depression. More specifically, my results indicated that women tend to experience severe depression at a higher rate than men.”



Reporting on Differences in Depression with a Chi-square test

Once you have conducted your t -test, paste the results into your portfolio. Report on the results using the template above.

Guided Research Project: Morality and the Heinz Dilemma

Module 5.3

Carry out a correlational research project examining moral decision-making

Throughout this chapter, you have explored how correlational research helps scientists understand relationships between variables. Now it is time to put this knowledge to work in a guided project.

For this project, we will return to the Heinz dilemma. Instead of describing people's moral decisions, however, we will examine what might predict those decisions. Specifically, we will use something called Moral Foundations Theory (e.g., Graham et al., 2013) to investigate whether differences in people's moral intuitions help explain their judgments in the Heinz dilemma.

The project will give you hands-on experience with the key elements of correlational research: developing theoretically driven hypotheses, analyzing relationships between variables, and interpreting results. The accompanying video for this project provides a step-by-step guide for what you need to do: [video link].

What We are Studying: Project Goals and Big Questions

You will complete three tasks for this project. First, you will generate research hypotheses. This will involve reading about Moral Foundations Theory and developing two specific hypotheses about how moral foundations might predict reactions to the Heinz dilemma. For each hypothesis, you will write approximately one paragraph explaining your rationale based on the theory.

Second, you will analyze the data. After examining the Qualtrics survey, you will download the SPSS data file and calculate scores for the five moral foundations subscales. You will then create a correlation matrix showing relationships between all five moral foundations and the moral acceptability ratings. Next, you will conduct *t*-tests to examine whether moral foundation scores differ between people who answered "yes" versus "no" to stealing the drug. For your hypothesized relationships, you will create appropriate figures, including a bar graph comparing moral foundation scores between yes/no groups.

Finally, you will write up the results. Your write-up should focus on the specific hypotheses you generated, following the examples from earlier in the chapter that show how to report statistical analyses.

Part 1: Frame Your Hypotheses

In Chapter 3, you conducted a descriptive study examining how people respond to the Heinz dilemma. You found that when people were asked a simple yes/no question about whether Heinz should steal the drug, a slight majority said he should not. You also found that when people were asked to rate how morally acceptable stealing would be, most found it somewhat acceptable.

In this project, you will explore what might predict people's responses. To answer this question, you will draw upon Moral Foundations Theory. Just as the Big Five personality traits provide a framework for understanding personality, Moral Foundations Theory suggests that moral judgments can be understood through five basic moral intuitions that are described in Box 5.4.

Brief Background Reading

This project begins where all research begins—by looking at existing theory. Take some time to read about Moral Foundations Theory. You can start at the www.moralfoundations.org website. Then, you can explore the peer reviewed literature on Google Scholar (e.g. Graham et al., 2013). During your reading, pay attention to how this theory has been used to understand people's moral judgments in different situations.

Develop Your Hypotheses

After familiarizing yourself with the Moral Foundations Theory, take a moment to develop two of your own hypotheses. Looking at the five moral foundations, which ones do you think might predict whether someone finds stealing the drug acceptable or unacceptable? Why?

Write down your predictions and your reasoning in your portfolio. Specifically, one hypothesis should relate to the question of why people find Heinz's decision morally acceptable or unacceptable. The second hypothesis should relate to the yes/no question asking whether Heinz should have stolen the drug. For example, are people who say that Heinz should steal the drug

more likely to have higher Authority scores compared to those who think he shouldn't steal the drug?

The Moral Foundations Theory

Care/Harm: This foundation focuses on our sensitivity to others' suffering and desire to protect the vulnerable. People high in care are compassionate, empathetic, and disturbed by cruelty. They prioritize alleviating suffering and providing care for those in need.

Fairness/Cheating: This foundation relates to justice, rights, and proportional treatment. People high in fairness value equality, reciprocity, and impartiality. They're sensitive to cheating, discrimination, and injustice, with strong reactions to those who violate these principles.

Loyalty/Betrayal: This foundation emphasizes obligations to in-groups like family, community, or nation. People high in loyalty value group cohesion, patriotism, and self-sacrifice for collective good. They strongly disapprove of betrayal and those who abandon their group commitments.

Authority/Subversion: This foundation relates to tradition, hierarchy, and leadership respect. People high in authority value social order, deference to legitimate authorities, and role fulfillment. They're concerned with maintaining institutions and respecting established hierarchies.

Sanctity/Degradation: This foundation involves physical and spiritual purity concerns. People high in sanctity are guided by disgust toward "unnatural" behaviors and value restraint, cleanliness, and treating certain things as sacred. They're concerned with avoiding contamination.

Liberty/Oppression: (added in later research) This foundation centers on sensitivity to freedom restrictions and domination resistance. People high in liberty value individual autonomy and are alert to signs of control or oppression. They're motivated to resist perceived tyranny and value personal freedom.

Box 5.4. Moral foundations theory.

Part 2: Design, Materials, and Methods

Now that you have hypotheses, let's examine how to test them. The first step is selecting the measures to assess each construct.

Moral Foundations Questionnaire

To measure moral foundations, we used the Moral Foundations Questionnaire (MFQ-30), which assesses how strongly people endorse each of the five moral foundations (Graham et al., 2008). The MFQ-30 contains 30 items total, with six items measuring each foundation. For instance, an item that measures the Care/Harm foundation asks people to rate how relevant "whether or not someone suffered emotionally" is when deciding if something is right or wrong. All items are answered on a 1 to 5 scale, although the answer labels vary across the measure.

Heinz Dilemma

We presented participants with the same moral dilemma and follow up questions used in Chapter 3. The questions asked whether Heinz should steal the drug (yes/no) and how morally acceptable stealing would be (rated from 1-7).

Data Quality

The Moral Foundations Questionnaire (MFQ-30) includes a few items that check to see if participants are paying attention. For example, one item asks participants how important "whether or not someone was good at math" is when deciding between right and wrong. To the items embedded within the MFQ, we added an additional attention check. Chapters 10, 11, and 12 cover data quality in-depth, but as an optional exercise, you can identify which participants failed the attention checks and remove them from the data analyses after performing the steps below.

Accessing study materials

To see how the measures were implemented online, download the Qualtrics survey file from the OSF project page: <https://osf.io/a8kev/>. Find the folder named "Ch. 5 – Correlational Research" and download the "RITC_SURVEY_CH05_HeinzDilemma.qsf" file. Import this file to your Qualtrics account and explore its structure.

Notice how the survey is organized into blocks, how we used matrix-style questions to present the MFQ-30 items, and how we randomly determined the order people received the MFQ-30 and

Heinz dilemma in (you can see this within the survey flow). The instructional video for this assignment walks you through the key features of this survey design in more detail.

Data collection

After programming the survey, we gathered data from 200 Connect participants. Each person was paid \$1.00 for their time, and the study took about 7 minutes to complete. Once we launched the project, data collection completed in under an hour.

Part 3: Analyze the Data

To analyze the data, download the SPSS file from the OSF page. In the “Ch. 5 – Correlational Research” folder find the file named “RITC_DATA_CH05_HeinzDilemma.sav” and download it.

Once you have the file open, calculate the scores for each moral foundation subscale. HOW TO Box 5.5 provides instructions or you can watch the video for this project. The MFQ-30 consists of five moral foundations with six items each. You need to average the items to create a single score for each foundation.

Once you have created a score for each subscale, create a correlation matrix to show the correlations between all five moral foundation subscales and the moral acceptability judgments. Then, conduct *t*-tests to see whether people who thought it was acceptable for Heinz to steal the drug differ from those who thought it unacceptable. These analyses will tell you whether your hypotheses about the moral foundations and the Heinz dilemma were supported.

Create a Figure

After the *t*-tests, create a bar graph that compares moral foundation scores between the yes and no groups. You can use either HOW TO Box 5.3 or the video for this assignment to guide you through this process. Make sure you put the yes/no groups on the x-axis and the moral foundation score on the y-axis. Also make sure your graph includes error bars (± 1 standard error). Label both axes and give the graph a title. You will include this figure with the write up of your results.

HOW TO: Analyze Data for the Heinz Dilemma Correlational Study

Download and open the dataset

- Navigate to the OSF page (<https://osf.io/a8kev/>)
- Download the "Heinz Dilemma - Correlational Project" SPSS file (.sav)
- Open SPSS and load the data file (File > Open > Data)

Calculate moral foundation scores

- Click "Transform" in the top menu, then select "Compute Variable"
- For each moral foundation, create an average score from its six items.
- For example, to create the "Harm" variable write "Harm" in the "Target Variable:" box. Then in "Numeric Expression" enter: MEAN(Harm1, Harm2, Harm3, Harm4, Harm5, Harm6)
- Click "OK"
- Repeat this process for the other foundations (Fairness, Loyalty, Authority, Sanctity)

Examine correlations with moral acceptability

- Click "Analyze" in the top menu, then select "Correlate > Bivariate"
- Move all five moral foundation scores and the "Acceptability" variable to the box
- Ensure "Pearson" is selected under "Correlation Coefficients"
- Click "OK" to produce the correlation matrix
- Identify which moral foundations correlate significantly with moral acceptability

Compare groups based on the yes/no decision

- Click "Analyze" in the top menu, then select "Compare Means > Independent Samples T-Test"
- Move the moral foundation scores to the "Test Variable(s)" box
- Move the "Steal" variable to the "Grouping Variable" box
- Click "Define Groups" and enter "1" for "Group 1" (Yes) and "2" for "Group 2" (No)
- Click "Continue" and then "OK"

Create a bar graph for your hypothesized relationship

- Click "Graphs" in the menu bar, then select "Chart Builder"
- In the gallery, select "Bar" chart and drag the simple bar chart to the canvas area
- Drag "Steal" to the x-axis and your foundation of interest to the y-axis
 - Make sure the "Steal" variable is listed as a "Nominal" measure otherwise SPSS will not recognize the categories for your chart
- Click the "Element Properties" button
- In the dialog box, select the "Error Bars" tab
- Check "Display error bars" and select "1 Standard Error" from the dropdown
- Click "Apply" and then "OK" to create your graph with error bars
- Repeat these steps for each of the Moral Foundations variables you hypothesized a difference for

Box 5.5. Instructions for analyzing the correlational Heinz dilemma.

Part 4: Interpret the Findings

How do the results compare to your hypotheses? Were your predictions supported? What surprised you about the findings? Why do you think some moral foundations predicted reactions to the dilemma while others didn't?

Write up the results relating to your hypotheses about moral foundations and acceptability ratings using the style below:

"I conducted a Pearson correlation analysis to examine the relationship between [your predicted moral foundation] and judgments of moral acceptability in the Heinz dilemma. There was a [significant/non-significant] [positive/negative] correlation between [foundation name] and moral acceptability ratings, $r(198) = [\text{value}]$, $p = [\text{value}]$."

Add one sentence explaining what your result means in plain language. Include your figure and a brief statement about whether your hypotheses were supported.

Group Differences

For your hypothesis about moral foundations and yes/no decisions, write up your findings with the following structure:

"I conducted an independent samples t -test to compare [foundation name] scores between participants who said Heinz should steal the drug and those who said he should not. There was a [significant/non-significant] difference in scores for yes ($M = [\text{value}]$, $SD = [\text{value}]$) and no ($M = [\text{value}]$, $SD = [\text{value}]$) groups; $t(198) = [\text{value}]$, $p = [\text{value}]$."

Add one sentence explaining what this means in plain language.



Part 5: Report the Relationship Between Moral Foundations and People's Decisions

For your portfolio, present your two hypotheses about moral foundations and people's judgments in the Heinz dilemma, with one paragraph explaining the rationale for each hypothesis.

Include your analyses and the write up from above. Present the correlation matrix showing relationships between moral foundations and acceptability ratings, followed by the *t*-test results comparing moral foundation scores between yes/no groups, and then the bar graph showing group differences for your effect.

Write up your results, with one paragraph each for the correlation and *t*-test results. End with a brief interpretation of what your findings mean.

Preview

Designing Your Own Correlational Study

Module 5.4

Apply your knowledge by designing, conducting, analyzing, and reporting your own correlational study

Now that you have worked through an example, you are ready to investigate your own correlational research question.

Throughout this book, you have explored several different domains of research. You have measured personality traits, examined clinical variables like anxiety and depression, and investigated stages of moral development. Now it is time to use what you have learned to investigate a question that interests you.

Your correlational study could build on any of the measures explored previously. For instance, you might wonder how personality traits relate to mental health: are more conscientious people less likely to experience anxiety? Or you might be curious about how moral foundations connect to other aspects of behavior: do people who score high on the care/harm foundation show more empathy in everyday situations?

You could also venture into new territory. Many students are curious about patterns they have observed in their own lives or questions about human behavior that intrigue them. For example: Do people who spend more time on social media report feeling more socially connected or more isolated? Are students who exercise regularly less stressed during exam periods? Do people who maintain structured daily routines sleep better? Is procrastination related to perfectionism?

You have experience finding validated measurement tools—from the TIPI for assessing personality to the GAD-7 for anxiety and the Moral Foundations Questionnaire for moral intuitions. You can use these measures, find new measures in the instrument databases we discussed, or create your own measure following the process outlined in Chapter 4 for working with AI. Either way, this project presents an opportunity to pull together several of the things you have learned to investigate a question of your choosing.

Step 1: Craft Your Question and Study Design

Using Qualtrics, design a short survey to collect data for a correlational research project. Keep it simple—measuring two or three variables is plenty. The best way to begin is to take the Moral Foundations study in Qualtrics and modify it. Include clear instructions for participants and organize your measurements into blocks.

The first block should include instructions and a welcome message. After that, you should measure one variable per block, placing all the items that are part of your measurement instrument into their own block. The last block should include basic demographic questions.

Step 2: Collect Data

For this project, you should aim to collect data from at least 100 participants. These could be students in your school (if your instructor sets up a class data collection system through a platform like Sona), friends and family (using the anonymous survey link from Qualtrics), or CloudResearch Connect participants.

Step 3: Analyze Your Results

Once you have the data, analyze it using the statistical tools you learned about in this chapter. For relationships between continuous variables, use correlations. For comparing groups on continuous measures, use *t*-tests. It is uncommon to examine relationships between categorical variables for this type of project. However, if this fits the question you have chosen, use chi-square tests. Create appropriate visualizations using the techniques you practiced in the chapter.

Step 4: Interpret and Share What You Found

The Appendix for Part I provides instructions for reporting research. Follow those guidelines and add your results to your portfolio. Include your research question and why you chose it, how you measured your variables, your statistical findings, a figure showing your main result, and your interpretation of what the results mean.



Stop and Discuss!

Before starting your project, discuss these questions with your class.

1. What research questions interest you and why?
2. What challenges do you anticipate in measuring your variables?
3. How will you recruit participants for your study?
4. What type of statistical analysis will best answer your research question?

Remember, the goal isn't just to find a statistically significant result; it is to learn something about human behavior through systematic investigation.

Summary

In this chapter, you explored correlational research. You learned how to analyze correlations, interpret their strength, and understand when relationships are statistically significant. You also saw how behavioral scientists assess relationships between different types of variables: continuous-continuous associations are measured with Pearson's r , categorical-continuous association with t -tests, and categorical-categorical associations with a chi-square.

The hands-on projects have given you practical experience with correlational analyses. By exploring relationships between anxiety, depression, and demographic variables, you had the chance to work with a real dataset. The Heinz dilemma project further demonstrated how psychological theories like Moral Foundations Theory can explain individual differences in moral judgments.

The goal of correlational research, as should be clear by now, is to identify patterns and make predictions. When behavioral scientists find that two variables are correlated, they know that one variable can predict the other. However, correlations alone cannot tell researchers whether one variable causes change in the other.

This limitation brings us to the next chapter, where we will explore how researchers address questions of causality in correlational research. Chapter 6 will introduce you to advanced

techniques for investigating cause and effect relationships when experiments are not possible. You will learn about statistical control, multiple regression, and longitudinal designs. These are common methods that help researchers build stronger evidence for causal relationships while still acknowledging the inherent limitations of correlational approaches.

By combining what you have learned about correlational research with these advanced techniques of causal inference, you will develop a more sophisticated understanding of how behavioral scientists investigate relationships between variables and draw meaningful conclusions about human behavior.

Preview

Chapter 6: Causal Inference

In the previous chapter, you learned how correlations reveal relationships between variables. You saw that when two variables are correlated, knowing someone's score on one variable helps to predict their score on the other. But predicting something isn't the same as knowing what caused it.

Just because two variables are correlated doesn't mean that changes in one variable cause changes in the other. When behavioral scientists say, "A causes B," they mean that manipulating variable A will create changes in variable B through some underlying mechanism.

Understanding these mechanisms is the key to changing behavior. If anxiety causes depression, for instance, then treating anxiety might prevent depression. But, if anxiety and depression are merely correlated because both are caused by some other factor, then treating anxiety might not affect depression.

Behavioral scientists encounter tricky questions about causation all the time. For example, many studies have found correlations between increased social media use and higher rates of depression and anxiety among teenagers. This research has led to calls for limiting teens' access to social media. However, the causal relationship is complex (e.g., Valkenburg, 2022). Does social media use cause mental health problems? Or do teens who struggle with mental health use social media more? Perhaps both variables are influenced by other factors like social isolation, family dynamics, or personality.

The stakes of answering these questions are high. Parents and schools invest millions of dollars in programs limiting teenagers' screen time and tech companies face calls for regulation. If the relationship between social media and mental health is not causal, these interventions will waste resources and fail to address teen mental health.

There are two reasons correlational research struggles to establish causation. The first is the directionality problem. When two variables are correlated, it is often unclear which variable influences the other. The second problem is the third-variable problem; two variables might be correlated because both are affected by something else—a third variable.

In this chapter, we will explore how behavioral scientists move beyond simple correlations to build a case for causation when experiments are not an option. Throughout the chapter, you will apply the techniques of causal inference in a series of guided analyses with the same clinical dataset you used in past chapters.

In Module 6.1, we will dig into the third-variable problem. You will learn how to use covariates in statistical analyses to examine relationships while holding potential confounding variables constant.

Module 6.2 tackles the directionality problem by introducing longitudinal research. You will learn how measuring variables at multiple points in time helps researchers establish temporal precedence. Knowing this allows researchers to argue in favor of a causal relationship.

Finally, Module 6.3 will give you the opportunity to design a correlational study that incorporates techniques of causal inference. You will develop hypotheses about causal relationships, plan appropriate statistical controls, and consider how these designs can strengthen your causal claims. By the end of the chapter, you will understand how behavioral scientists build evidence for causality in correlational research, and you will have practical experience applying these techniques to real data.

Chapter 6 Outline: Causal Inference

Module 6.1: Controlling for Third Variables



Laboratory

Explore how statistical controls help researchers rule out third variables and make stronger claims about causal relationships in data

- Explore the limits of correlations, including the third variable problem
- Learn what it means to statistically control for third variables
- Conduct your own analyses to control for third variables

Module 6.2: The Directionality of Cause and Effect



Laboratory

Examine the directionality problem by learning about temporal precedence and how longitudinal designs can strengthen causal interpretations.

- Read about the importance of temporal precedence in establishing causality
- Explore how longitudinal research can establish temporal precedence
- Conduct time-lag correlational analyses with a longitudinal dataset

Module 6.3: Designing Your Own Causal Inference Study



Laboratory

Design a study that controls for confounds and makes a compelling case for causality.

- Develop a research question and identify potential third variables
- Design a study to test your ideas and rule out third variables
- Collect, analyze, interpret, and report your own data

Controlling for Third Variables

Module 6.1

Explore how statistical controls help researchers rule out third variables and make stronger claims about causal relationships in data

When behavioral scientists discover a correlation between two variables, it raises an important question: Does this relationship reflect a causal connection or is it explained by other factors?

Internal validity refers to how confident researchers can be that the effect they have observed is due to a cause-and-effect relationship, rather than the result of unmeasured or confounding variable.

Internal validity is often at the heart of debates about how to apply the results of research to problems in daily life, as in this example.

Decades of research have found a correlation between playing violent video games and aggressive behavior in adolescents (e.g., Anderson & Bushman, 2001). Based on this research, some groups have called for regulating video games. But, the issue of internal validity raises important questions. First, do video games lead teenagers to behave aggressively or, perhaps, do teenagers who are prone to aggressive behavior prefer to play violent games? This question is referred to as the **directionality problem** (see Figure 6.1).

Whenever researchers observe a correlation between two variables, the causal effect could go in either direction. The mere existence of a correlation between playing violent video games and aggression does not allow us to disentangle which variable is the cause and which is the effect. Indeed, research on this topic articulates two hypotheses. The *socialization hypothesis* states that playing violent video games increases aggression over time. Meanwhile, the *selection hypothesis* states that aggressive teenagers are more likely to play violent video games over time.

The Directionality Problem

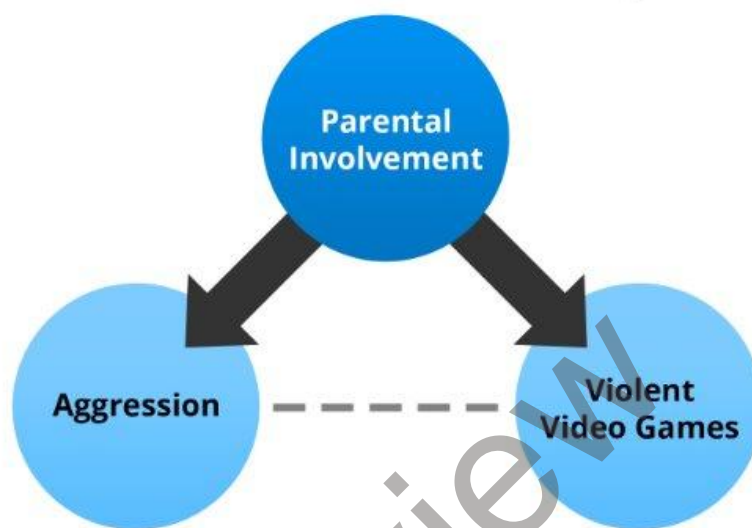


When two variables are correlated,
we can't tell which direction causation might flow.

Figure 6.1. The directionality problem with violent video games and violent behavior.

Separate from the directionality problem there is another reason that the correlation between violent video games and aggressive behavior cannot establish causation: the **third-variable problem**. Simply stated, the third variable problem is the idea that something besides violent video games and aggressive behavior, something like a lack of parental supervision, could cause the correlation (see Figure 6.2). If a third variable exists, then there is no causal relationship between violent video games and aggressive behavior—they are only correlated because both are caused by lack of parental involvement. And, of course, if playing violent video games does not cause aggressive behavior, then regulating violent games will do little to reduce aggressive behavior (Przybylski & Weinstein, 2019).

The Third Variable Problem: Violent Video Games Explains Aggression-Parental Involvement Relationship



The apparent relationship between aggression and violent video games (dashed line) may be explained by their common relationship with low parental involvement

Figure 6.2. A third variable may explain why violent video games and violent behavior are correlated.

Why Control Matters: Strengthening Internal Validity

There are a few things researchers can do to strengthen the internal validity of correlational findings. One thing is to systematically identify and control for potential third variables. Controlling third variables helps rule out alternative explanations and builds a stronger case for causality. The second thing researchers can do is conduct longitudinal research. Longitudinal research gathers data from participants at several points over time, helping establish which

variable is more likely to cause the other. We will examine how researchers control for third variables in this module before turning to longitudinal research in the next module.

A Thought Experiment: How to Control for Third Variables

One way to understand the logic of controlling for third variables is to engage in a thought experiment. Let's take the example of violent video games and adolescent aggression.

Suppose you suspect that parental involvement explains this relationship. To rule out that possibility, you could design a study that includes only teenagers whose parents are highly involved in their lives. In other words, every participant in the study would have parents who are maximally involved in their lives, as measured by a validated scale. Now, if you find that violent video games predict higher aggression among this group of teenagers, you would have good reason to believe the relationship is *not* due to a lack of parental supervision. However, by making parental involvement into a constant, so it is the same for everyone in the study, you have also learned something important: parental involvement doesn't explain the relationship.

While it is rarely practical to control variables this way in real life, researchers can achieve the same goal statistically. Instead of selecting only participants who are equal on some variable, such as parental involvement, statistical techniques allow researchers to ask: What *would* the relationship between two variables look like if everyone were equal on a third variable? That is the principle behind the analysis you are about to see.

Controlling Third Variables with ANCOVA: Marital Status, Depression, and Age.

Let's examine how this works using a finding from the clinical dataset you worked with in past chapters. The finding involves the relationship between marital status and depression. In the dataset, single people reported higher depression ($M = 7.4$) than those who were married ($M = 5.8$). This difference is statistically significant ($t = 2.5, p < .05$), indicating an association between depression and marital status (Figure 6.3). You can reproduce these results by following the instructions in Box 6.1

HOW TO: Instructions for conducting the t-test on depression and marital status.

Open the Dataset

- Open SPSS and load the "RITC_DATA_CH06_ClinicalStudy.sav" dataset from the Chapter 6 folder on OSF

Filter the Data to Include Only Single and Married Participants

- Click on Data \square Select Cases
- Select "If condition is satisfied" and click on "If..."
- In the formula box, enter: $\text{Mar} = 1 \mid \text{Mar} = 6$. This selects cases where marital status is either 1 [Single] or 6 [Married].
- Click "Continue" and then "OK"

Run an Independent Samples t-test

- Click on Analyze > Compare Means \square Independent-Samples T Test
- Move Depression SCORES to the "Test Variable(s)" box
- Move mar (MaritalStatus) to the "Grouping Variable" box
- Click "Define Groups" and set groups (1 = Single, 6 = Married)
- Click OK to run the analysis

Create a Bar Chart

- Click on Graphs > Chart Builder > Bar
- Move mar (MaritalStatus) to the "X Axis" box
- Move depression SCORES to the "Y axis" box, select "Mean" as the statistic
- Click "Display Error bars" and then select standard error. Change the value to 1
- Click "OK" to create the chart

Remove Empty Categories in Chart Editor

- Double-click on the chart to open the Chart Editor
- Click on any of the category labels on the x-axis to select them
- Double click on a label again and select "Categories..." from the pop up menu
- In the Categories dialog box, select each empty category (e.g., divorced) and then click the red "X" to delete it.
- Remove all categories except for Married and Single
- Click "Apply" and then "Close" to update the chart

Box 6.1. Instructions for conducting the *t*-test on depression and marital status.



Figure 6.3. Association between marital status and depression. Married people are less depressed than single people.

At face value, you might interpret this result to mean that something about being married protects people from depression. In other words, you might assign a **causal explanation**.

After observing any association, however, it is important to think about potential third variables. In the case of marriage and depression, what third variables might explain this effect? Before reading on, see if you can come up with a plausible third variable explanation.



Stop and Discuss!

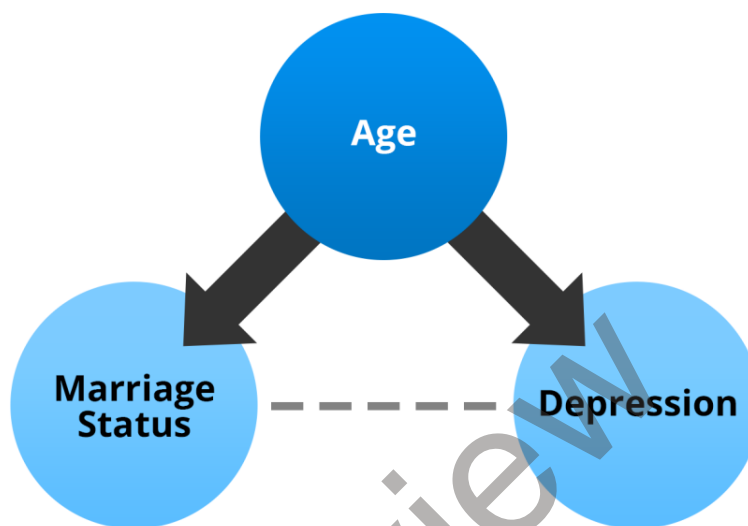
- 1) Look at Figure 6.1. What does it mean to say that there is an apparent causal relationship between marriage and depression?
- 2) Thinking of the relationship between marriage and depression, what third variable(s) could explain the relationship between marriage and depression?

Here is a potential explanation for the relationship between marriage and reduced depression that does not invoke marriage as a cause of lower depression. You may remember from Chapter 3 that the National Institutes of Mental Health survey found older adults tend to be less depressed than younger adults (see Figure 3.2). We also saw something similar in the previous chapter where age was negatively correlated with depression (see Figure 5.5).

At the same time, we know age is correlated with marital status. People start life single and some later get married. Indeed, the average age of married people in the analysis shown in Figure 6.3 was 44.5 years old while the average age of single people was 36.5 years old, an eight-year difference. Given that people are less depressed as they get older *and* they get married as they get older, it is possible married people are less depressed because they are older, not because they are married.

The relationship in Figure 6.4 identifies the problem: age is a plausible third variable that might explain why married people report lower depression than single people. In this situation, age is referred to as a **covariate**—a third variable that influences both marital status and depression. If that is the case, we cannot be confident that marriage is the cause of the observed difference in depression. So how do we figure out whether age is driving the effect? The answer lies in the logic we applied to the video game example.

Third Variable Effect: Age Explains Marriage-Depression Relationship



The apparent relationship between marriage and depression (dashed line) may be explained by their common relationship with age.

Figure 6.4. A demonstration of the third-variable problem. Age may explain why married people are less depressed than those who are single.

Let's imagine a controlled version of a study in which researchers hold age. If researchers conducted a study and sampled only people who are exactly 40 years old, then they could compare depression scores between married and single people. Because everyone in the study would be the same age, any differences in depression could not be due to age. If the researchers found that married 40-year-olds were less depressed than single 40-year-olds, they would know age does not account for differences in depression. If, on the other hand, they found that married 40-year-olds were just as depressed as single 40-year-olds, it would suggest that age—and not marriage—explains the original relationship.

Figure 6.5 shows a visual representation of the hypothetical design. The key to this design is understanding that age has been made into a **constant** rather than a variable. *Everyone in the study is the same age.* As you may remember from Chapter 3, constants do not change across people in a study. By making age a constant, researchers can remove it as an explanation for any differences observed within the study.

Examining Marriage-Depression When Age is Controlled

By examining people who are all the same age, we can see if marriage truly predicts differences in depression.

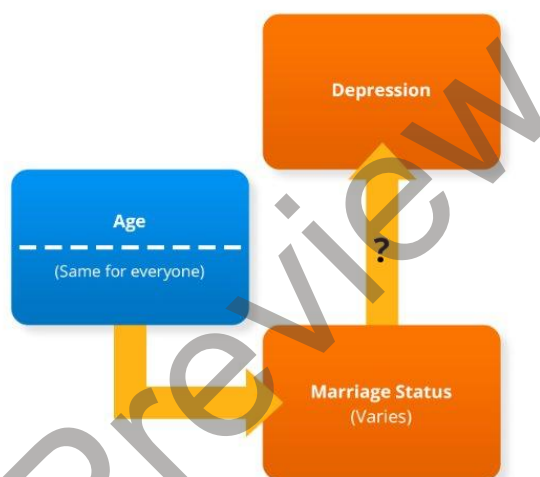


Figure 6.5. Controlling for age involves making age into a constant and examining whether the association between anxiety and depression remains intact.

Activity 6.1 – Statistically Controlling Third Variables 📊

Designing studies the way you were asked to imagine in the thought experiment—where researchers select only people of the same age—is almost never done in practice. That is because recruiting people who are exactly the same on age or exactly the same on any other characteristic would be extremely difficult. Instead, researchers use statistical techniques that mathematically transform a variable into a constant, allowing them to examine relationships *as if* everyone in the sample were the same age.

When age is added as a covariate, the statistical software calculates what the results would look like *if* everyone in the sample were the same age. Let's look at how this works in an analysis.

The analysis required to examine whether married people are less depressed than single people while statistically controlling for age is called an ANCOVA—Analysis of Covariance. An ANCOVA is appropriate when one variable is categorical (marital status) and the other is continuous (depression). ANCOVA will tell us what the mean difference between the married and single groups would be if everyone in the sample was the same age and whether that adjusted difference is statistically significant.

You can follow the steps in HOW TO Box 6.2 to run the ANCOVA and create the bar chart required for this activity or you can watch the accompanying video: [video link].

When age is controlled for in the analysis, the difference in depression between married ($M = 6.10$) and single ($M = 6.97$) people shrinks and is no longer statistically significant ($p > .05$).

Figure 6.6 shows what depression scores would look like if everyone in the sample were exactly forty-one and a half years old (the average age of the sample). Notice there is still a difference in depression between groups, but the difference is much smaller than in the original analysis. The smaller difference between groups and loss of statistical significance suggests that age, rather than marital status itself, might better explain why married people are less depressed.

Specifically, *our results suggest that married people in our sample are less depressed because they are older.*

This example illustrates a broader point about correlational research. While behavioral scientists cannot establish causation through correlational methods, controlling for third variables allows them test hypotheses about whether specific third variables are playing a role. Sometimes, controlling for a third variable reveals that what appeared to be a meaningful relationship is better explained by other factors (as in the example above). At other times, the original relationship remains strong even after controlling for third variables.

HOW TO: Control for Third Variables Using ANCOVA in SPSS

Open the Dataset

- Open SPSS and load the "RITC_DATA_CH06_ClinicalStudy.sav" dataset if it is not already open

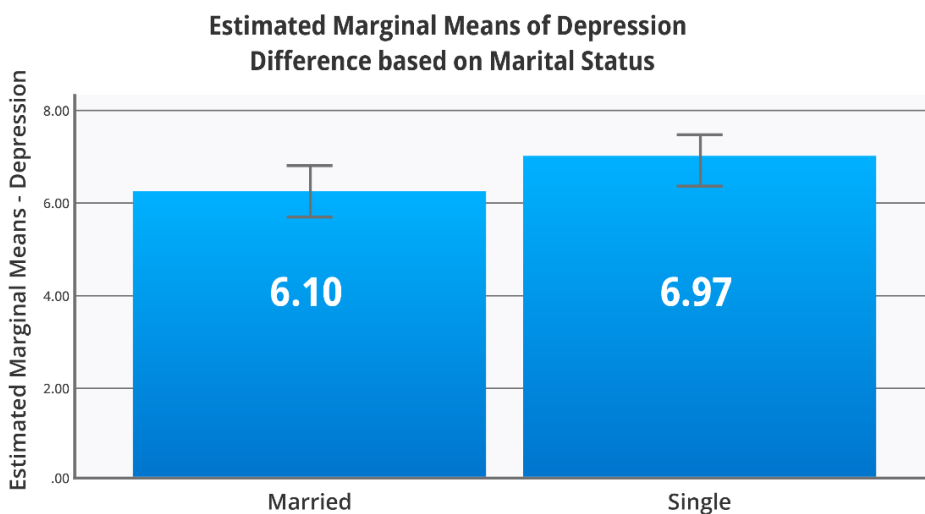
Filter the Data to Include Only Single and Married Participants

- Click on Data Select Cases
- Select "If condition is satisfied" and click on "If..."
- In the formula box, enter: Mar = 1 | Mar = 6. This selects cases where marital status is either 1 [Single] or 6 [Married].
- Click "Continue" and then "OK"

Run an ANCOVA analysis

- Click on Analyze in the top menu
- Select "General Linear Model" --> "Univariate"
- Move Depression SCORES into the Dependent Variable box
- Move mar (MaritalStatus) into the "Fixed Factor(s)" box
- Move "Age" into the "Covariate(s) box"
- Next, select "EM Means..."
 - Choose "mar" (Marital Status) and move it into the "Display Means for:" box
 - Select "Compare main effects" and then within the "Confidence interval adjustment:" box use the dropdown to select "Bonferroni"
- Before running the analysis, create a bar chart by selecting "Plots..."
 - Move "mar" from the "Factors:" box over to the "Horizontal Axis": box
 - Select "Add"
 - Change the chart type to "Bar Chart"
 - Click the box to "Include Error Bars" and select "Standard Error"
 - Change the standard error multiplier to "1." Then click "Continue"
- Click OK to run the analysis

Box. 6.2. How to conduct an ANCOVA and create a bar chart.



What is your relationship or marital status?

Covariates appearing in the model are evaluated at the following values:
How old are you? = 41.51 Error bars: +/- 1 SE

Tests of Between-Subjects Effects

Dependent Variable: DEPRESSION SCORES

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	655.575 ^a	2	327.788	10.016	<.001
Intercept	3404.101	1	3404.101	104.013	<.001
age	435.289	1	435.289	13.300	<.001
mar	63.320	1	63.320	1.935	.165
Error	12665.617	387	32.728		
Total	29411.000	390			
Corrected Total	13321.192	389			

p > .05

a. R Squared = .049 (Adjusted R Squared = .044)

Figure 6.6. Marital status is no longer a statistically significant predictor of depression ($p = .17$) when age is statistically made into a constant.

Each time the original relationship withstands the scrutiny of controlling for another third variable, it increases the researcher’s confidence that there is meaningful effect. For example,

behavioral scientists studying whether violent video games cause aggressive behavior in adolescents have controlled for many third variables, including parental supervision, socioeconomic status, age and developmental stage, competitiveness, mental health, the influence of peers, and several others (e.g., Adachi & Willoughby, 2011; Anderson & Bushman, 2002; Greitemeyer & Mügge, 2014; Markey & Markey, 2010). Over the years, the association between violent video games and aggressive behavior has remained. And, the ability to control for third variables in this way is a fundamental part of the research process because it helps build a stronger case for causality in correlational research.



Describing the Relationship Between Marital Status and Depression after Controlling for Age

Once you have conducted the analysis above, paste the ANCOVA output to your portfolio. Also paste the bar graph of the original difference in depression between married and single people and the bar graph where age was controlled for.

Write a few sentences describing what the output shows about the relationship between age, depression, and marital status. Is depression associated with marital status after age is controlled for? Why?

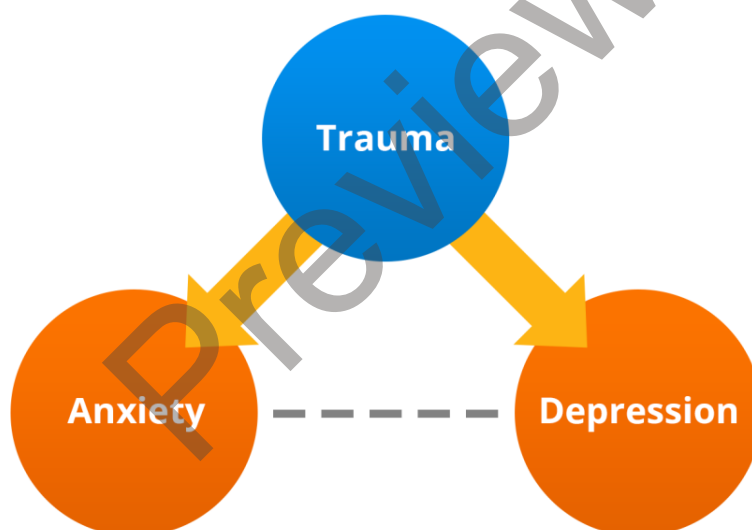
Controlling for Third Variables with Regression

In the previous section, you saw how controlling for age led to a better understanding of the relationship between marital status (a categorical variable) and depression (a continuous variable). The same principle applies when examining correlations between two continuous variables, but a different statistical analysis is used.

Let's return to the correlation between anxiety and depression ($r = .82$), which are both continuous variables. While it might be tempting to conclude that anxiety causes depression, or perhaps that depression causes anxiety, it is possible that some other factor causes both anxiety and depression.

For instance, one plausible explanation for the anxiety-depression relationship involves trauma (Figure 6.7). People who live through traumatic events often experience distress and develop both anxiety and depression as a result. Common traumatic events include divorce, job loss, serious financial problems, academic failure, relationship breakups, family conflict, moving away from home, serious illness, being bullied or socially excluded, and the end of a close friendship. If trauma causes both anxiety and depression, then the correlation between depression and anxiety might arise because both caused by a history of trauma.

Third Variable Effect: Trauma Explains Anxiety-Depression Relationship



The apparent relationship between anxiety and depression (dashed line) may be explained by their common relationship with trauma.

Figure 6.7. A third variable such as trauma may explain why anxiety and depression are correlated.

To test this idea, we could use a statistical technique called multiple regression. Multiple regression is like a correlational analysis but with an additional capability: it allows researchers to examine how two variables correlate while controlling for other, third variables. Just as controlling for age in the previous example showed what the marriage-depression relationship would look like if everyone were the same age, regression allows researchers to see what the relationship between anxiety and depression would look like *if* everyone in the sample had experienced the same level of trauma. In particular, we can examine what the correlation between anxiety and depression would be if we only recruited people who had no trauma. This means the analysis would calculate the anxiety-depression correlation when trauma is not playing a role.¹ It does so with something called a partial correlation.

Partial Correlations Explained

When researchers conduct a regression analysis one key piece of information they get is called a **partial correlation**. A partial correlation tells how strongly two variables are related after statistically holding other variables constant. Sometimes the partial correlation becomes much smaller than the original correlation and may no longer be statistically significant. This would indicate that the variable the researcher is controlling for explains a lot of why two variables were related. Other times, the partial correlation remains close to the original correlation, suggesting that the variable the researcher is controlling for does not explain much about why the two variables are related.

Looking at the size of the original correlation compared to the partial correlation helps explain whether the covariates are playing an important role in the relationship. If controlling for a third variable (trauma) causes the original correlation (anxiety–depression) to become much weaker or disappear entirely, it suggests the third variable might be the real reason the other two variables are related. If the correlation remains strong, even after controlling for other variables, it suggests

¹ A technical note about what happens when a variable is held statistically constant in regression or ANCOVA is warranted. The default value for the covariate is set to the average of the sample. This means if we control for trauma, the covariate will be evaluated at the average trauma score of the sample. For example age in the previous example, set to 41 years old, the sample average. However, the covariate can be held constant at any value, including zero. The value does not change the outcome of the analysis. In other words, the partial correlation value will remain the same independent of whether the covariate is set to the mean or to zero. The only thing that changes is the intercept, something that we do not discuss in this book. For our purposes, we will either talk about setting the covariate to the mean of the sample or to zero, depending on the conceptual clarity of each case.

there is a relationship that exists independently of the third variables that were controlled for. Let's look at an example.

Research Activity 6.2: Multiple Regression in Action

In this activity, you will see how the anxiety-depression relationship changes when controlling for people's past traumatic experiences. To run the regression analysis, you can either follow the steps in the HOW TO Box 6.3 or watch the video for this activity: [video link].

Preview

HOW TO: Control for Third Variables Using Regression in SPSS

Open the Dataset

- Open SPSS and load the "RITC_DATA_CH06_ClinicalStudy.sav" dataset if it is not already open

Run a linear regression analysis

- Click on "Analyze" in the top menu
- Select "Regression" --> "Linear"
- Move Depression SCORES into the Dependent Variable box
- Move both the predictor variable "Anxiety" and the control variable "Trauma" into the "Independent (s)" box

Request partial correlations

- In the Linear Regression dialog box, click on "Statistics"
- Check "Part and partial correlations"
- Click "Continue"
- Back in the main dialog box, click "OK" to run the analysis

Interpret your results

- In the "Coefficients" table, look at the significance values (Sig.) for each predictor
- If your predictor variable (e.g., Anxiety) remains significant ($p < .05$) even with the control variable in the model, this suggests a robust relationship
- Look at the "Correlations" section of the output which shows:
 - Zero-order correlations (original correlation without controlling for other variables)
 - Partial correlations (relationship after controlling for other variables)
- Compare the zero-order correlation with the partial correlation to see how much the relationship changes after controlling for the third variable

Box 6.3. Instructions for how to conduct a multiple regression analysis.

As a reminder, the original correlation between anxiety and depression was $r = .82$. When we conduct the regression analysis and receive the partial correlations—what the correlation would be if everyone in the sample had no trauma-related distress—we see the anxiety-depression correlation is $.76$, which remains statistically significant (Figure 6.8). This tells us two important things.

First, trauma explains some of the relationship between anxiety and depression. You can see this in the reduced correlation, from .82 to .76. However, the partial correlation remains quite strong (.76), which tells you the second important thing: anxiety and depression have a robust relationship that exists even after controlling for trauma. This means that even if everyone in the sample had no trauma-related distress, there would still be a strong correlation between anxiety and depression.

Model		Unstandardized Coefficients		Standardized Coefficients		Coefficients ^a		Zero-order	Correlation	
		B	Std. Error	Beta	t	Sig.	Partial		Part	
1	(Constant)	1.138	.212		5.369					
	ANXIETY SCORES	.830	.031	.747	26.575	<.001		.826	.758	.638
	TRAUMA SCORES	.527	.098	.151	5.383	<.001		.540	.229	.129

a. Dependent Variable: DEPRESSION SCORES

Figure 6.8. The relationship between anxiety and depression remains statistically significant even after statistically making trauma into a constant

Reporting Results: What Does the Regression Show?

Here is an example of how to report the results of this analysis. You can use this as a template for reporting your own results.

“I conducted a multiple regression analysis to examine the relationship between anxiety and depression while controlling for trauma-related distress. The zero-order correlation between anxiety and depression was strong, $r = .82$, $p < .001$. After statistically controlling for trauma-related distress, the partial correlation between anxiety and depression decreased slightly but remained significant, $r_p = .76$, $p < .001$. These results suggest that while trauma explains part of the association, a strong relationship between anxiety and depression persists even after accounting for trauma.”

Controlling for Multiple Third Variables at Once

Of course, trauma-related distress is not the only variable that might explain the relationship between anxiety and depression. Poor sleep could play a role: people who sleep poorly often experience both anxiety and depression. Age might also be important. As we saw earlier, age relates to both depression and anxiety. Finally, socioeconomic factors like income and education might contribute, since they affect both anxiety and depression.

Multiple regression allows researchers to control for *all* these variables simultaneously. What does this mean? Rather than controlling for each variable one at a time, it is possible to statistically calculate what the correlation between anxiety and depression would look like if everyone in the sample had the same trauma score AND the same sleep quality AND the same age AND the same income AND the same education level (Figure 6.9). It is just like the examples we have seen, but now the analysis holds several variables constant at once. Let's give it a try.

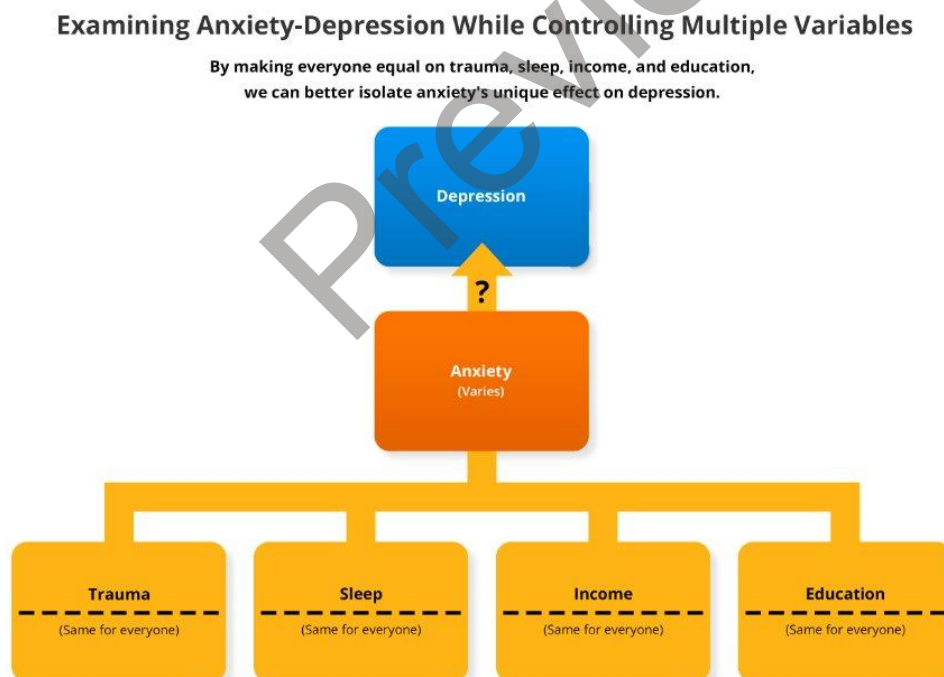


Figure 6.9. A statistical model examining the correlation between anxiety and depression while holding trauma, sleep, income, and education constant.

Research Activity 6.3: Anxiety and Depression, Controlling for Trauma, Sleep, and More



To conduct this analysis, return to the Clinical dataset. Then, you can either follow the steps on HOW TO Box 6.4 or watch the video for this activity: [video link].

When you control for all variables in the analysis simultaneously, the partial correlation between anxiety and depression drops to .656 (Figure 6.10). This drop is larger than when controlling for trauma alone—the partial correlation then was .76. The reduction in the correlation indicates that the additional variables in the analysis explain some of why anxiety and depression are related.

HOW TO: Control for Multiple Third Variables Using Regression in SPSS

Open the Dataset

- Open SPSS and load the "RITC_DATA_CH06_ClinicalStudy.sav" dataset if it is not already open

Run a multiple regression analysis

- Click on "Analyze" in the top menu
- Select "Regression > Linear"
- Move "Depression" into the "Dependent" box
- Move all variables into the "Independent(s)" box:
 - Anxiety (main predictor of interest)
 - Trauma (third variable #1)
 - Sleep (third variable #2)
 - Income (third variable #3)
 - Education (third variable #4)
- Request partial correlations
 - In the Linear Regression dialog box, click on "Statistics"
 - Check "Part and partial correlations"
 - Click "Continue"
- Back in the main dialog box, click "OK" to run the analysis

Interpret your results

- In the "Coefficients" table, look at the significance values (Sig.) for each predictor • Focus on the partial correlation value for Anxiety (your main predictor of interest)
- Compare this partial correlation to the zero-order correlation to see how much the relationship changes after controlling for all these variables simultaneously

Box 6.4. Instructions for how to control multiple third variables using regression.

Looking at the analysis reveals that several control variables have their own relationship with depression. Sleep quality, for instance, shows a significant partial correlation of .419 with depression, suggesting that poor sleep is independently related to depression even after controlling for everything else, including trauma and anxiety. Trauma has a smaller but still significant relationship with depression (partial correlation = .157). Income, however, is not significant once the other factors are controlled for.

Model	Coefficients ^a					Correlations			
	Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.	Zero-order	Partial.	Part	
1	(Constant)	-1.311	.608		-2.157	.031			
	Anxiety	.652	.033	.586	19.825	<.001	.825	.656	.431
	Trauma	.327	.091	.094	3.616	<.001	.540	.157	.079
	Sleep	.285	.027	.297	10.508	<.001	.711	.419	.228
	What is the highest level of education you have completed?	-0.122	.105	-.027	-1.163	.245	-.061	-.051	-.025
	What was your total household income last year?	-0.057	.033	-.041	-1.719	.086	-.168	-.075	-.037

a. Dependent Variable: DEPRESSION SCORES

Figure 6.10. The relationship between anxiety and depression remains statistically significant even after statistically making trauma, sleep, income, and education into constants.

Yet even after accounting for all the control variables in this analysis—essentially asking what the anxiety-depression relationship would look like if everyone were identical on all the control characteristics—the partial correlation of .656 remains high and statistically significant. This suggests that while trauma, sleep, income, and education together explain part of the relationship between anxiety and depression, there is something robust about the relationship between anxiety and depression that exists independently of these other factors.



Reporting What these Controls Reveal about Anxiety and Depression

Once you have conducted the analysis, paste the regression output to your portfolio. The output should include the original correlation (zero-order) and the partial correlation.

Write a few sentences describing what the output shows about the relationship between depression, anxiety, and trauma-related distress. Interpret the partial correlation in your own words. Is trauma-related distress the likely cause of depression and anxiety?

Why Statistical Control Strengthens Causal Claims

Throughout this module, we have explored how behavioral scientists address the third-variable problem in correlational research. Whenever researchers discover a relationship between two variables—like marriage and depression or anxiety and depression—they must consider whether other factors might explain the relationship. By statistically controlling for potential third variables, it is possible to build stronger evidence for causal claims.

We have examined how researchers do this with two examples. First, we found that the relationship between marital status and depression became non-significant after controlling for age. This suggests that age, rather than marriage itself, might explain why married people in our sample reported lower levels of depression. Second, we discovered that while trauma, sleep quality, income, and education explain some of the relationship between anxiety and depression, a strong association remains even after controlling for these factors.

The statistical techniques introduced thus far are powerful tools. They allow researchers to explore the potential causes of the associations they observe. Would married and single people still differ in depression if they were the same age? Would anxiety and depression still be related if everyone had the same level of trauma or the same quality of sleep?

The answers to these questions bring researchers closer to understanding causal relationships, even within correlational designs. When a relationship disappears after controlling for a third variable (as with marriage and depression), it suggests the original relationship may have been spurious. When a relationship remains robust after controlling for several alternative explanations (as with anxiety and depression), however, researchers gain more confidence in the findings' importance—even though they still cannot definitively claim causation.

Ruling out third variables is essential when conducting research. Whether researchers are studying the effects of social media on teen mental health, the relationship between video games

and aggressive behavior, or the connection between exercise and mood, statistical control techniques help separate genuine relationships from those better explained by other factors. In the next module, we will build on these methods by exploring how to address another key challenge in establishing causality: determining which variable comes first.

Preview

The Directionality of Cause and Effect

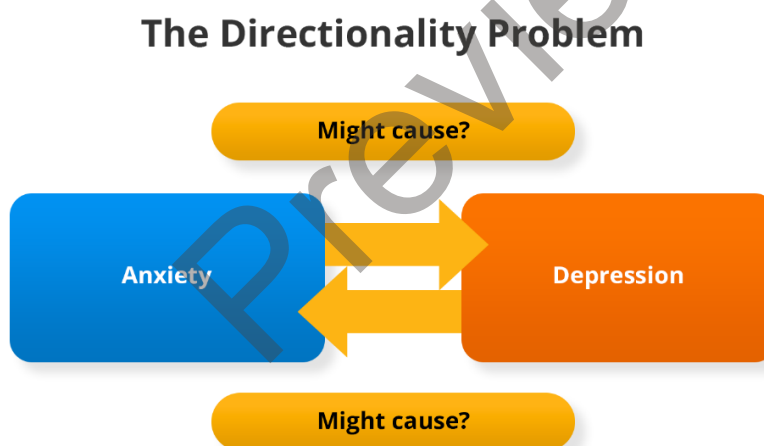
Module 6.2

Examine the directionality problem by learning about temporal precedence and how longitudinal designs can strengthen causal interpretations.

The Directionality Problem: Which Comes First?

In the previous section, you learned how researchers address the problem of third variables. But correlational research also faces the directionality problem. When two variables are correlated, it is often unclear which variable causes changes in the other. Let's explore how researchers tackle directionality by returning to the relationship between anxiety and depression.

As we have discussed, there is a strong positive correlation between anxiety and depression. Yet the problem of directionality asks: does anxiety lead to depression, or does depression lead to anxiety (Figure 6.11)?



When two variables are correlated, we can't tell from the correlation alone whether one causes the other, or which direction causation might flow.

Figure 6.11 An illustration of the directionality problem.

Think about how this might play out in the life of a friend. Maybe your friend experienced anxiety about an upcoming exam, which led to difficulty sleeping and concentrating. As their

anxiety persisted, they started feeling hopeless about their academic performance, lost interest in activities they usually enjoy, and developed symptoms of depression. In this case, anxiety preceded and potentially contributed to depression.

But the opposite is equally plausible. Perhaps your friend first experienced depression, feeling unmotivated and struggling to keep up with coursework. As the assignments piled up, they became increasingly anxious about falling behind, developing symptoms of anxiety that were not present before. In this case, depression led to anxiety.

The question of directionality has practical application as it affects what interventions might be effective. If anxiety leads to depression, then preventing or treating anxiety early might help prevent depression from developing. But, if the causal direction runs the other way, then early intervention for anxiety won't do much to prevent depression.

Behavioral scientists face the question of directionality whenever they find a correlation between two variables. For example, does social media use lead to loneliness or does loneliness lead to social media use (Keles et al., 2020)? Do violent video games increase aggression or do people who are naturally aggressive choose violent games (Anderson & Bushman, 2001)? Does exercise improve mood or are people in better moods more likely to exercise (Hyde et al., 2011)? A correlation cannot answer these questions alone.

Why Temporal Precedence Matters

For one thing to cause another, the cause must come before the effect. This simple idea—called **temporal precedence**—is a requirement of causality. If researchers want to say depression causes anxiety, they need to show that increases in depression *precede increases in* anxiety. Similarly, if anxiety causes depression, increases in anxiety must precede increases in depression.

A simple correlation between anxiety and depression measured at one point in time does not tell researchers anything about temporal precedence. All that a correlation reveals is that two variables tend to occur together. It says nothing about which variable developed first or whether changes in one variable preceded changes in another.

To establish temporal precedence, variables need to be measured at different points in time. To do that, behavioral scientists often turn to longitudinal designs.

Establishing Temporal Precedence with Longitudinal Research

Longitudinal research examines how variables relate to each other over time. Rather than measuring variables just once, longitudinal studies measure the same variables multiple times over a period of days, weeks, months, or even years. This approach makes it possible to untangle how the variables influence each other.

Let's return to the example of anxiety and depression. Instead of measuring these variables once, imagine a researcher measured them twice—once at the beginning of the semester (Time 1) and again at the end of the semester (Time 2). This simple change in design opens powerful possibilities for understanding how anxiety and depression relate to each other.

To demonstrate the possibilities, we gathered data from 480 Connect participants over a one-year period. Each participant completed measures of anxiety and depression at two timepoints, one year apart. With this dataset, we are interested in whether depression at Time 1 predicts increased anxiety at Time 2 and vice versa: does anxiety at Time 1 predict depression at Time 2?

Correlations between two different variables across time are called **cross-lag correlations**, while a correlation between two variables at one point in time is called a **cross-sectional correlation**. (see Figure 6.12). When we examine the cross-lag correlation between depression at Time 1 and Anxiety at Time 2, we see that depression at Time 1 *predicts* anxiety one year later. As Figure 6.13 shows, in a simple correlational analysis there is a strong correlation between depression at Time 1 and anxiety at Time 2.

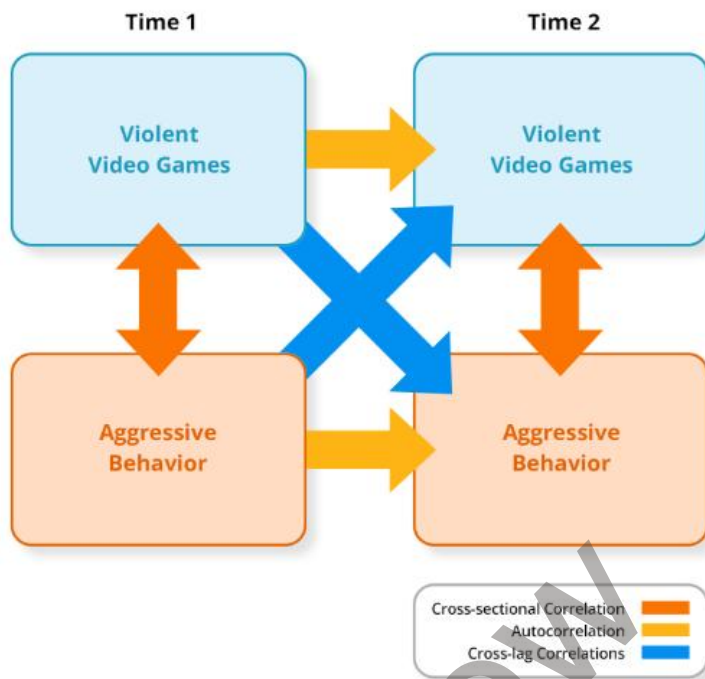


Figure 6.12. Longitudinal designs yield a variety of correlations.

Correlations

		t1depression	t2anxiety
t1depression	Pearson Correlation	1	.669**
	Sig. (2-tailed)		<.001
	N	479	479
t2anxiety	Pearson Correlation	.669**	1
	Sig. (2-tailed)	<.001	
	N	476	480

** Correlation is significant at the 0.01 level (2-tailed).

Cross-lag correlation between depression at Time 1 and anxiety one year later.

p. < .05

Figure 6.13. An example of a cross-lag correlation: depression measured at Time 1 predicts anxiety one year later.

The cross-lag correlation above establishes that depression scores predict anxiety scores one year later. This is important because prediction is a requirement of causation: for one variable to cause another it must predict that variable. But while we have established longitudinal prediction, the correlation alone does not establish temporal precedence. This is because people who are depressed at Time 1 also have higher levels of anxiety at Time 1 than people who are not depressed. If the researchers found that people who are depressed at Time 1 have higher anxiety at Time 2 how would they be able to tell if it is because of depression at Time 1 or anxiety at Time 1? In other words, people who are depressed at Time 1 are also anxious, which may be why their depression at Time 1 is correlated with anxiety at Time 2.

Answering the question above requires statistical controls. By controlling for anxiety at Time 1, we can examine whether higher levels of depression predict future increases in anxiety *among people who started with no anxiety* (Figure 6.14).

Let's once again consider a simplified version of the study to clarify the statistical analysis. Imagine that in September of 2024, we recruited participants who varied in their levels of depression but everyone reported *no* symptoms of anxiety. In other words, every participant had an anxiety score of zero as measured by the GAD-7. One year later, in September of 2025, we brought the same participants back and measured their levels of depression and anxiety. Suppose we found that many of the participants who had no anxiety in 2024 now showed signs of anxiety. We could then conduct a *cross-lag correlation* to examine whether depression in 2024 predicted the rise in anxiety one year later.

If we found that people who were more depressed in 2024 were more likely to develop anxiety in 2025 under these conditions, we would have strong evidence of *temporal precedence*. In this scenario, depression was present before anxiety developed. Importantly, if we statistically *hold anxiety constant* in 2024—meaning we compare people who all started out with no anxiety—and still find that depression predicts anxiety in 2025, then we can conclude that depression both *precedes* and *predicts* anxiety. That is, people who began with similar levels of anxiety but differed in depression show different anxiety outcomes a year later, with those higher in initial depression experiencing greater increases in anxiety. This pattern would provide strong evidence that depression precedes the development of future anxiety.

This is what the regression analysis we will conduct below accomplishes. It calculates what the cross-lag correlation between depression in 2024 and anxiety in 2025 would be *if* everyone had no anxiety in 2024.

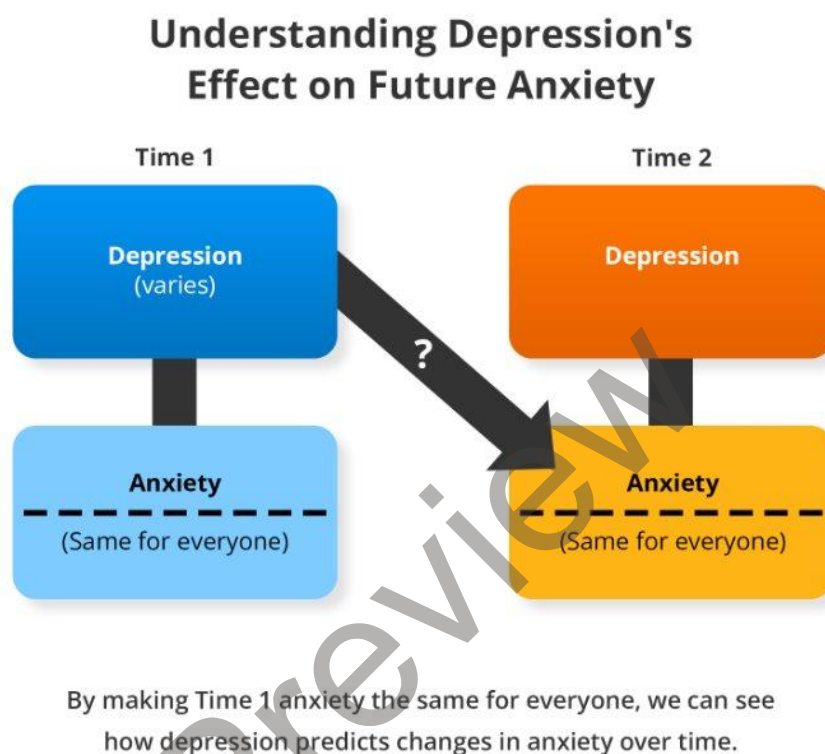


Figure 6.14. Cross-lag correlation between depression at Time 1 and anxiety one year later, while holding anxiety constant at Time 1.

Research Activity 6.4: Correlation Between Anxiety and Depression One Year Later: Time Lagged Correlations 📊

You can download the dataset for this activity from the *Research in the Cloud* OSF page: <https://osf.io/a8kev/>. In the folder labeled “Chapter 6 – Causal Inference” you will find a file titled “RITC_DATA_CH06_Longitudinal.sav” Download this file and open it in SPSS.

To conduct the analysis, we will turn to regression. In SPSS, select "Regression," → Linear and put Time 2 anxiety in the dependent variable box. Then, in the Independent(s) box add Time 1 depression and Time 1 anxiety. This will calculate the cross-lag correlation between depression

at Time 1 and anxiety one year later, while acting as if everyone had the same level of anxiety at Time 1.

As displayed in Figure 6.15, the association between depression at Time 1 and anxiety at Time 2 remains statistically significant.

Coefficients ^a

Model		Standardized Coefficients		Standardized Coefficients Beta	t	Sig.	Zero-order	Correlation	
		B	Std. Error					Partial	Part
1	(Constant)	1.394	.249		5.596	<.001			
	t1anxiety	.645	.067	.612	9.566	<.001	.730	.403	.300
	t1depression	.119	.056	.135	2.116	.035	.668	.097	.066

a. Dependent Variable: t2anxiety

Anxiety is being predicted from depression one year later

p. < .05, a significant effect

Figure 6.15. Cross-lag correlation between Anxiety at Time 1 and Depression one year later, having controlled for initial levels of anxiety

What do these results say about the relationship between depression and anxiety? First, they reveal that depression at Time 1 predicted anxiety one year later. Second, and more importantly, the results show this relationship held even after controlling for initial anxiety. This means among people who started with no anxiety at Time 1, those with higher depression scores showed a larger increase in anxiety over the following year than those with lower depression scores. Such a pattern suggests that depression's ability to predict anxiety one year later cannot be explained simply by pre-existing anxiety levels. In other words, we have established temporal precedence for depression.

Overall, this example illustrates how longitudinal research can help scientists better understand the complex relationships between psychological variables. By measuring variables at multiple time points and controlling for people's initial levels of something like anxiety, researchers can build a stronger case for how one variable might influence another over time.

Reporting the Results of a Cross-Lagged Regression

Here is an example of how to report the results of this analysis. You can use this report as a template for your own results.

“I conducted a cross-lagged regression analysis to examine whether depression at Time 1 (September 2024) predicted anxiety at Time 2 (September 2025), controlling for baseline anxiety. The results indicated that Time 1 depression significantly predicted increases in anxiety one year later, even after accounting for initial levels of anxiety, $B = .12$, $t(478) = 2.1$, $p < .05$. This suggests that higher depression at baseline was associated with higher anxiety at follow-up, independent of initial anxiety levels.”



Describing the Longitudinal Relationship Between Anxiety and Depression

Once you have conducted the analysis, paste the regression output to your portfolio. The output should include the original cross-lag correlation between depression and anxiety and the partial correlation with anxiety controlled for at Time 1. Write a few sentences describing what the output shows about the relationship between depression, anxiety, and temporal precedence. Explain the finding in your own words. Finally, conduct an analysis establishing the temporal precedence of anxiety relative to depression. Is there evidence for bidirectional causality?

Combining Approaches: Multiple Controls in Longitudinal Research

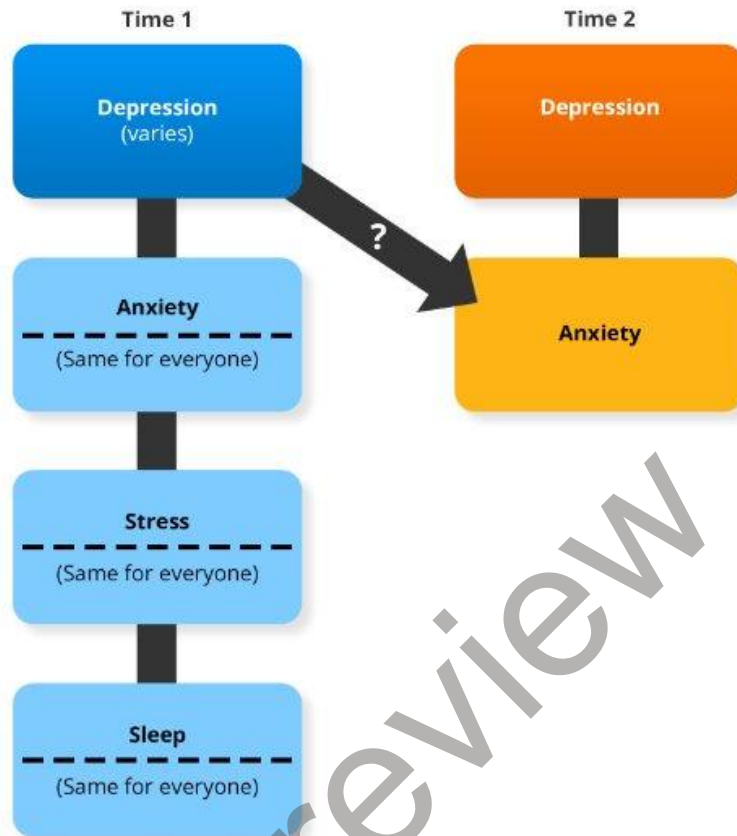
At the beginning of this chapter, we discussed two challenges in establishing causality: the third-variable problem and the directionality problem. You have now seen how statistical controls can address the third-variable problem, and how longitudinal designs can address the directionality problem. But the most compelling evidence for causality in correlational research comes from combining these approaches. By controlling for multiple third variables in a longitudinal study, researchers can build a compelling case about the causal relationships between variables. For instance, scientists studying social media's effects on teen mental health might control for pre-existing conditions, family factors, and other variables while tracking changes over time to make a stronger case for causality (see Przbylski and Weinstien, 2019).

Let's see how this plays out with our example of anxiety and depression. While a longitudinal study helps researchers understand which variable comes first, third variables can still play a role. Even when depression predicts future anxiety after controlling for initial anxiety levels—as we found in the previous example—third variables might still be causing both. Remember the third variables discussed earlier—trauma, sleep quality, and income? These could still influence both depression and future anxiety.

To rule out these explanations, it is possible to combine approaches. Just as the previous study controlled for initial anxiety levels, it could also control for other important variables measured at Time 1. For instance, we could statistically control for people's initial levels of traumatic stress, making everyone's stress level the same at the start. We could also do the same with sleep quality and social support. By controlling for all these variables at Time 1, we not only demonstrate the temporal precedence of depression relative to anxiety, but also rule out that both depression and anxiety are being caused by other third variable covariates (Figure 6.16).

This approach is particularly valuable because psychological variables rarely operate in isolation. Mental health, like most aspects of human behavior, involves complex relationships between many variables. By controlling for multiple factors at once, researchers can better understand the unique role that each variable plays in predicting future outcomes.

Controlling for Multiple Variables Over Time



By controlling for multiple variables at Time 1, we can better isolate depression's unique effect on future anxiety.

Figure 6.16. Predicting anxiety at Time 2 from depression at Time 1 while controlling for multiple variables at Time 1.

Designing Your Own Causal Inference Study

Module 6.3

Design a study that controls for confounds and makes a compelling case for causality.

In the previous modules, we discussed how behavioral scientists address the third-variable problem through statistical controls and the directionality problem through longitudinal research. Now it's time to apply these techniques in your own correlational study that incorporates the methods of causal inference.

From Correlation to Causal Inference: Your Research Project

Throughout this book, you have been building research skills step by step. In Chapter 4, you learned to measure psychological constructs and to find and develop scale instruments. In Chapter 5 you examined correlations between variables. Now you will take your research to the next level by addressing potential third variables that might explain the correlations you observe.

This project gives you the opportunity to design and conduct a correlational study that controls for a relevant third variable. You will collect real data on Connect, analyze it using the techniques you have learned, and interpret your findings by comparing the original uncontrolled effect to the effect that is observed after controlling for a covariate.

Designing and Conducting a Causal Inference Study

For this project, you will design and conduct a correlational study that examines a potential causal relationship between two variables while controlling for at least one plausible third variable.

Step 1: Choose your research question

Start by selecting a research question that interests you. Your question should focus on a potential causal relationship between two variables. You have several options for approaching this project.

First, you might build on your Chapter 5 project. If you conducted a correlational study in the previous chapter, you can expand it here by identifying and measuring potential third variables that might explain the relationship you found. This approach allows you to deepen your investigation of a topic you have already started exploring.

Second, you can draw from measures we have explored in previous chapters. These include personality traits from Chapter 1; moral foundations from Chapter 3; clinical variables such as anxiety, depression, or sleep quality from Chapters 5 and 6; or variables that can predict responses to moral dilemmas like the Heinz dilemma. Using established measures simplifies your design process and connects your work to the broader research literature.

Third, you might create your own measures. Using techniques from Chapter 4, particularly AI-assisted scale development, you can create measures for constructs not covered by existing scales. This approach gives you more flexibility to investigate what interests you.

When forming your research question, focus on how variables relate to each other when accounting for other factors. Your question should examine relationships that are both theoretically interesting and practically measurable using a measurement instrument of between five and ten questions. For example, you might ask: is openness to experience associated with people's willingness to engage with opposing political viewpoints when controlling for political orientation? Or is the care/harm moral foundation related to donation intentions toward different charitable causes when controlling for agreeableness or empathy? Any of these questions, or hundreds of others, will work.

Step 2: Identify Potential Third Variables

Once you select a research question, identify at least one potential third variable that might explain the relationship between your main variables. For each third variable, consider why it might be related to your predictor variable, why it might be related to your outcome variable, and how it could potentially explain the relationship between the variables.

The process of identifying third variables requires both creativity and knowledge of your topic. Think about what factors might influence both of your main variables. Consider demographic

characteristics, environmental factors, or psychological traits that could give rise to both variables you are interested in.

Step 3: Design Your Study

After you have identified your variables, plan your study. Your plan should specify how you will operationalize each variable. For instance, which scale or questions will you use to measure your predictor variable, outcome variable, and third variables? Be specific. If you are creating new measures using AI, document your process following the guidelines from Chapter 4.

Next, consider participant recruitment. As in previous projects, you can recruit participants from your university's participant pool using SONA, send your study to friends and family, or use Connect. You should aim for around 100 participants, if possible.

Using the skills you developed in previous chapters, create your survey in Qualtrics. Organize your survey into clear, logical blocks that guide participants through the survey experience. Include attention checks within your survey for data quality.

You do not need to build your Qualtrics study from scratch. Take an existing correlational study from the previous chapter and add another block for the control variable. You can swap the variables in that study for the ones you will use in this one.

Step 4: Collect and analyze Your Data

If you are collecting data from Connect, you can follow the step-by-step instructions from Chapter 3. Once your data collection is complete, download your data for analysis. Using SPSS or another statistical package, begin your analysis by examining the bivariate correlations between your predictor and outcome variable. Note the strength, direction, and statistical significance of these correlations.

Next, control for your identified third variables using multiple regression or ANCOVA, depending on your variable types. For continuous outcome variables, multiple regression is typically appropriate. For categorical predictors and continuous outcomes, ANCOVA would be

the method of choice. The statistical techniques we practiced in this chapter will guide your analysis.

After running these analyses, examine how the relationship between your main variables changes when controlling for third variables. Does the correlation become stronger, weaker, or stay about the same? Does it remain statistically significant, or does the significance disappear once you control for third variables? These changes provide clues about the nature of the relationship between your variables.

Step 5: Interpret Your Findings

Based on your analysis, interpret what your findings suggest about potential causal relationships. If controlling for third variables substantially weakened or eliminated the relationship between your main variables, this suggests that the third variables might explain much of the original correlation. The apparent relationship between your main variables might be spurious rather than causal.

Conversely, if the relationship remains strong despite controlling for plausible third variables, this provides more confidence in a causal connection, though it still does not prove causation. There might be other unmeasured variables that explain the relationship, or the direction of causality might run in the opposite direction from what you hypothesized.

Consider what other explanations might exist for the patterns you observed. What other variables might you want to control for in future research? How might experimental methods address some of the limitations of your correlational approach? This critical thinking about alternative explanations is a key aspect of scientific reasoning.

Writing it Up: Telling the Story of Your Study

After completing your study, prepare a 2-to-3-page research report that presents your project. Follow the instructions in the Part I Appendix for writing research reports. Your results section should report your findings, including both the simple correlations and the relationships after controlling for third variables.

In your discussion section, interpret your results in terms of potential causal relationships. Discuss what your findings suggest about whether one variable influences the other. Acknowledge limitations of your approach, such as the inability to manipulate variables experimentally or potential unmeasured third variables. Suggest directions for future research that might address these limitations.

Summary

Throughout this chapter, we have explored how behavioral scientists use sophisticated correlational techniques to build a case for cause-and-effect relationships. By statistically controlling for third variables, conducting longitudinal research to establish temporal precedence, or combining both approaches, researchers can build stronger evidence about how variables may be influencing each other in the moment or over time. For instance, when behavioral scientists find that depression predicts future anxiety even after controlling for initial anxiety, stress, sleep problems, and other variables, they are building a pattern that is consistent with depression being the cause of anxiety.

But correlational research, no matter how sophisticated, can never establish causation. There is always the possibility that some variable that was not measured or controlled for might explain the observed relationship. Because it is not possible to control for every possible third variable, researchers can never rule out all alternative explanations through correlational research alone.

This limitation is precisely why experiments are so valuable in behavioral research. In experimental studies, which we explore in the next chapter, researchers don't just measure and control for variables statistically—they manipulate variables to see what happens. They might randomly assign some participants to complete a stress-reduction program while others serve as a control group, then measure how this intervention affects both depression and anxiety. This kind of manipulation, combined with random assignment, provides the strongest possible evidence for causal relationships.

Yet this does not mean correlational research isn't valuable. Many important variables in behavioral science—like depression, anxiety, aggression, personality traits, or life experiences—cannot be easily manipulated in an experiment for either ethical or practical reasons, if they can be manipulated at all. In these cases, carefully designed correlational studies that combine

longitudinal measurement with appropriate controls provide the best window into understanding cause and effect. These studies help researchers develop theories about how psychological processes work, which can then be tested more rigorously through experimental methods, when possible.

While correlational research cannot definitively prove causation, the statistical control techniques you have learned in this chapter can substantially strengthen causal inference. By thoughtfully controlling for third variables, you can build more compelling evidence for potential causal relationships.

Preview

Chapter 7: Experimental Research

When you hear the word "experiment," you might think of a scientist in a white lab coat mixing chemicals or recording observations through a microscope. While some behavioral experiments take place in laboratories, the essence of an experiment isn't about location or equipment; it is about methodology. An experiment is a carefully controlled investigation where researchers manipulate one variable to observe its effect on another.

Within the behavioral sciences, descriptive research helps scientists understand a phenomenon, and correlational research identifies relationships between variables. But only experimental research can determine if one thing causes another. For example, a correlational study might show that anxiety and depression are related, but it cannot determine whether anxiety causes depression, whether depression causes anxiety, or if both are caused by something else entirely. Only experiments can answer causal questions, and in this chapter, you will learn how this is accomplished.

In Module 7.1, you will explore the fundamentals of experimental design. You will learn how experiments address the directionality and third-variable problems that limit correlational research. By understanding the key elements of manipulation, independent variables, dependent variables, and random assignment, you will discover why experiments are considered the gold standard for establishing causality.

Module 7.2 guides you through a hands-on project using the Heinz dilemma. You will learn about a common experimental manipulation known as perspective-taking and how it affects people's moral judgments. You will then manipulate perspective taking by randomly assigning participants to conditions, and you will analyze the resulting data with appropriate statistical tests. This module will give you hands-on experience with the steps of conducting a simple experiment.

In Module 7.3, you will explore variations on experimental design, focusing on repeated measures experiments. You will learn how within-subjects designs allow each participant to experience multiple conditions, enabling more powerful comparisons with fewer people.

Through a guided project examining how different consequences affect moral judgments, you will practice implementing counterbalancing and analyzing repeated measures data.

Finally, Module 7.4 introduces factorial designs that allow researchers to examine multiple variables simultaneously. You will discover how these complex experiments reveal not just main effects but also interactions—situations in which the effect of one variable depends on the level of another. Through a factorial study of perspective-taking and wealth, you will learn to design and analyze more complex experiments.

By working through each module, you will develop an understanding of what experiments are, how they establish causality, and the different ways they are used within the behavioral sciences. By the end of the chapter, you will have the knowledge and skills to design, conduct, and analyze your own experimental studies.

Preview

Chapter 7 Outline: Experimental Research

Module 7.1: How Experiments Establish Causality



Reading

Learn what defines an experiment and how manipulation, control, random assignment, and replication help researchers isolate causal effects.

- Explore the critical elements of an experiment and how they establish cause-and-effect relationships
- Read about common applications of experimental design
- Create your own experiment in Qualtrics

Module 7.2: Guided Project: Can Perspective-Taking Shift Moral Judgment?



Laboratory

Conduct an experimental study on perspective-taking and moral judgment

- Learn about perspective-taking and how it is manipulated in psychological research
- Examine the study materials and build the experiment in Qualtrics
- Analyze the data and report the results
- Design your own experimental study

Module 7.3: Variations on Experimental Design: Repeated Measures Experiments



Laboratory

Explore repeated measures designs by conducting a guided project on how different consequences influence moral decision-making.

- Read about repeated measures experimental designs
- Conduct a guided project examining how the consequences of an action shape people's moral judgments
- Conduct repeated measures analyses and report the results

Module 7.4: Beyond Simple Experiments: Factorial Designs



Laboratory

Investigate interactions by running a factorial experiment on perspective-taking and moral judgments.

- Explore why researchers sometimes manipulate more than one variable in an experiment.
- Conduct a factorial experiment to examine people's moral judgments
- Analyze the data and report the results

How Experiments Establish Causality

Module 7.1

Learn what defines an experiment and how manipulation, control, random assignment, and replication help researchers isolate causal effects.

What is an Experiment?

You are probably familiar with colloquial usage of the word experiment. It means to test, to try out, and to learn from experience. That's what scientists do when conducting experimental research. Unlike correlational studies, which measure variables and examine their association, experiments require researchers to act. To see how, consider this example.

Imagine you are participating in a study about social attitudes. The researcher hands you a survey and asks you to answer a simple question: "What are your views about abortion?" You consider the issue and mark your response. For the second question, you are once again asked about abortion but this time the item reads: "What do you think God's view is about abortion?" Just like the first question, the response options range from strongly oppose to strongly support.

Now, consider a slightly different scenario. You walk into the lab and are presented with two questions about abortion. But this time, the order is reversed. The question about God's views on abortion comes first, followed by the question about your own views. Would your answers be the same in this scenario as in the previous one? For many people, they are not.

The scenarios above are from an experiment conducted by psychologists Benjamin Converse and Nicholas Epley (2007). They wanted to test whether considering God's perspective *causes* people's position toward abortion to change. They found that people who are asked to think about God's views before their own view often express less support for abortion than people who share their personal opinion first. Although this experiment raises many interesting questions about religion and morality, we chose it to highlight how simple an experiment can be.

Key Elements of the Experimental Method

First, notice how the researchers created two groups of participants by manipulating something specific: the order of the questions (Figure 7.1). Some participants considered their personal views first, while others considered God's views first. This simple **manipulation** is an example

of an **independent variable**—the thing researchers deliberately change within an experiment to see what effect it has.

Experimental Design Flow

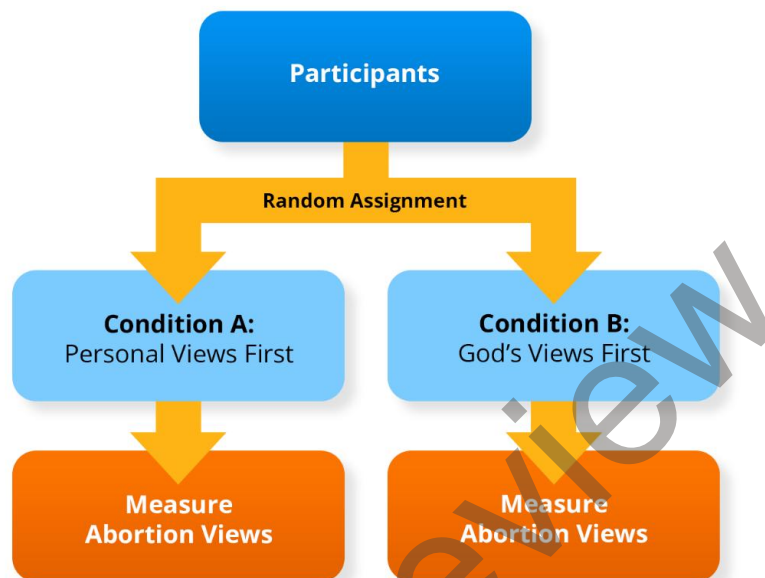


Figure 7.1. Experimental design of the God views study.

Second, consider what the researchers measured: people’s views about abortion. This is the **dependent variable** because researchers hypothesize that it *depends on*, or is affected by, the independent variable. By measuring people’s views about abortion after the manipulation, Converse and Epley were able to examine whether thinking about God's perspective first caused people to express less support for abortion compared to when they reported their own attitude first.

Finally, and most importantly, participants were assigned to groups within the study using **random assignment**. Random assignment is like flipping a coin: heads you go to condition A (asking about personal views first), tails you go to condition B (asking about God's views first).

Even though it is simple, random assignment is powerful. When people are randomly assigned to a condition in an experiment, any pre-existing differences—in things like age, education, religious beliefs, attitudes about abortion, and countless other factors—are equally distributed across conditions. In other words, random assignment solves the third-variable problem by ensuring that everything that could influence people's attitudes about abortion is neutralized as an explanation for behavior. Everything, that is, except for the experimental manipulation.

The Logic of Random Assignment

Random assignment is so important to experiments that it is worth analyzing in depth.

Let's imagine you are conducting a version of the study about God's view and abortion. You recruit 500 participants: 250 people who strongly support abortion and 250 who strongly oppose abortion. These are people's pre-existing views, before the study begins.

As each person enters the study, you flip a coin to determine whether they will see questions about God's view first (heads) or their own view (tails). When you flip the coin for the 250 pro-choice participants, each person has a 50-50 chance of being in either condition. The same is true for each pro-life participant.

It is important for each person to have an equal chance of being in either condition because their pre-existing views about abortion could, obviously, affect the results. If more pro-life people wound up in the God-first condition, you would not know whether to attribute that group's scores to the manipulation or their pre-existing views. The same thing is true in the opposite direction for pro-choice people. But, with random assignment, you do not have to worry about this problem. If you have enough people in your study, the coin flip ensures people with strong pro-life views and people with strong pro-choice views are distributed about equally between conditions.

And herein lies the magic of random assignment: the same logic that applies to people's pre-existing views about abortion applies to every other characteristic that might affect attitudes about abortion, too. Some participants might be influenced by their religious upbringing, others by personal experiences, and still others by political ideology. Random assignment ensures that all the factors you can think of, and the ones you can't, have an equal chance of appearing in either condition (Figure 7.2).

How Random Assignment Controls Third Variables

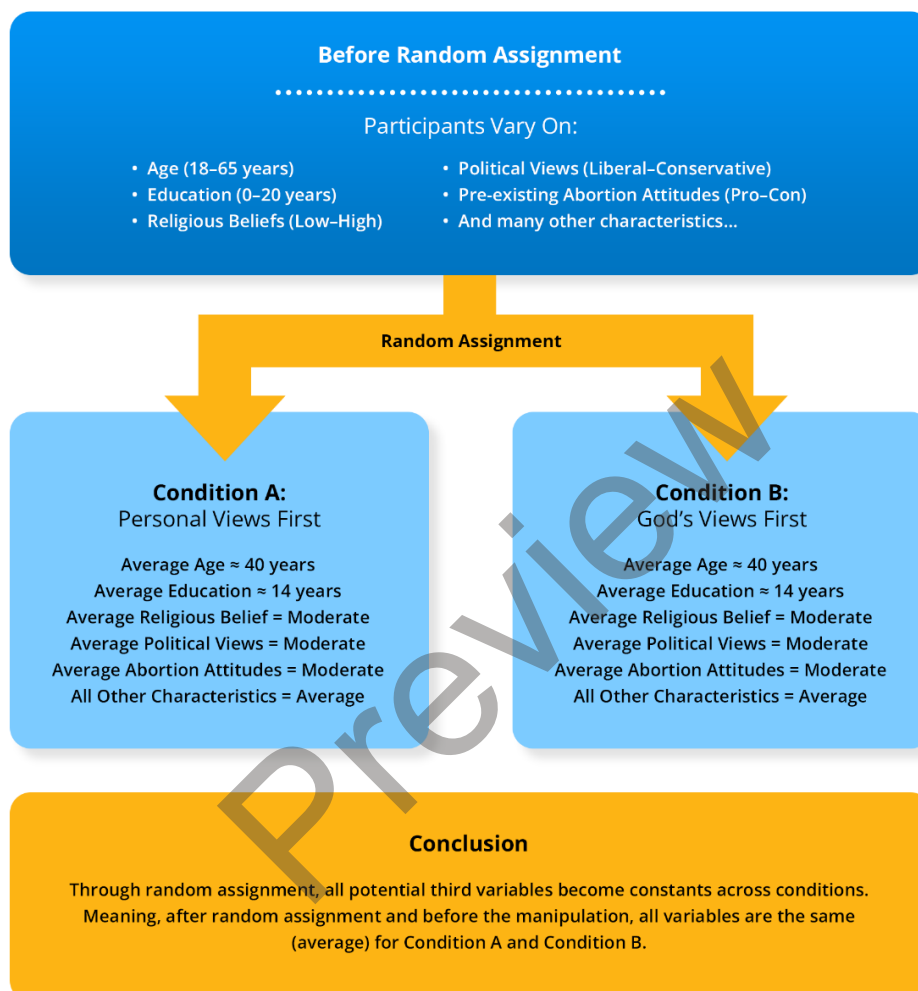


Figure 7.2. Random assignment distributes all variables equally across conditions. In effect, this makes all variables constants across conditions – the means of variables are equal across conditions.

Another way to think about random assignment is to realize that it turns every variable that might affect the results of your experiment into a constant across the two conditions. This means people in Condition A and Condition B will have about the same values, on average, across all potential

third variables. In fact, the only thing that differs between the two conditions for certain is the experimental manipulation, which means any differences you observe in the dependent variable must be caused by whether participants thought about God's views first or second.

Sample Size and Replication

Of course, random assignment is not perfect. If you flip a coin ten times, you might get seven heads and three tails just by chance. Similarly, if you only had ten participants, you might end up with more pro-choice people in one condition than the other. But as you recruit more people and your sample size increases, the odds of these imbalances decrease. With 100 participants, it would be rare for random assignment to create large differences between conditions on any variable. With 500 or more participants, such differences are virtually impossible. This is why sample size matters so much in experiments.

The larger the sample, the more confident researchers can be that random assignment has effectively controlled for all possible third variables. Most experiments aim for at least 50 participants per condition—enough that random assignment can work its magic. But even with large samples, any single experiment might get "unlucky" with its random assignment. This is why replication—conducting the same experiment multiple times—is important. Even if one study happens to have an uneven distribution of some important third variable, it is unlikely that multiple replications would have the same uneven distribution. When several studies show the same effect, researchers can be increasingly confident that the manipulation really causes a change in the dependent variable.

Experiments versus Correlational Studies

Think about how different the God's View study is from correlational research.

In a correlational study, researchers might ask whether people's religious views correlate with attitudes about abortion. The researchers would measure people's religious beliefs and attitudes about abortion. They might find that as people's religious belief increases, their support for abortion decreases. But, this correlation could not tell the researchers whether people's religious beliefs cause anti-abortion views, whether having anti-abortion views draws people to religion, or whether some other factor explains both.

The experimental approach, in contrast, manipulates which thoughts are salient when people express their views about abortion and controls for all possible alternative explanations. This is why experiments are considered the gold standard for establishing causation in behavioral science. By carefully controlling the research situation, manipulating specific variables, and holding others constant through random assignment, experiments allow researchers to draw strong conclusions about cause-and-effect relationships.

Where Experiments Appear in the Real-World

Experimental methods have a big impact on society. One of the most consequential applications is in clinical trials that test new drugs and medical treatments. Consider how these trials work.

Just as the God Views study randomly assigned participants to different conditions, a **clinical trial** randomly assigns patients to receive either the experimental treatment, such a new drug, or a **placebo**. A placebo is an inert pill that looks like real medication but lacks active ingredients. In a clinical trial, the placebo acts as a control group—a baseline to compare against the effect of the drug. Even though clinical trials often contain several groups where researchers can test different doses of the drug, at least one treatment group and the control group are essential. When participants are randomly assigned to these conditions, every factor that might affect health besides the drug—age, stress, mental health, exercise, diet—is equally distributed between groups.

Clinical trials also contain another critical feature: they are typically "**double-blind**." In a double-blind study neither the patients nor the researchers interacting with them know who received the drug versus the placebo. This prevents people's expectations from influencing the study's results.

If patients knew they were receiving the real drug, for instance, they might expect to feel better and report improved health or reduced anxiety even if the drug was not actually effective. In fact, there is an entire body of research on what is called the "placebo effect" where people report all kinds of changes in health when given a placebo. Evidence suggests the placebo effect is not just an illusion; people's expectations create real physiological changes (Price et al, 2008). A double-blind design also prevents researchers interacting with participants from consciously or

unconsciously treating people in one group differently than the other, potentially biasing the study's outcomes. For these reasons, a double-blind design is standard in clinical trials.

Experimental design also underlies progress in nearly every corner of society—from how farmers grow food to how businesses design websites, from policing strategies to political messaging. Any time someone wants to isolate the effect of one change—whether it's a new fertilizer, a redesigned app, or a public health campaign—they turn to the core logic of experiments: randomly assign people to groups, hold everything else constant, and observe what happens. These methods allow people to make causal claims with confidence, which is why they are so widely used.

In fact, once you start looking, you will see experiments everywhere. Companies test advertising strategies with what they call “A/B tests”. Governments evaluate new policies using randomized trials. Sports scientists test training routines. Even dating apps run experiments to see which profiles people are most likely to swipe on. Although the tools may look different in each setting, the underlying logic is the same. Experimental methods strip away noise and uncover whether one thing causes another or not.

Research Activity 7.1: Design an Experiment

Now that you understand the fundamentals of experimental research, let's examine how to create an experiment.

The God Views study provides a perfect example of how to implement an experiment online. It also illustrates a valuable idea: experiments do not have to be complex to be important. An experiment can consist of nothing more than two sentences presented in a different order. As long as the manipulation is sound and participants are randomly assigned to conditions, you can create an interesting experiment. Let's look at how to do this using Qualtrics.

The video for this activity will show you how to create the God views experiment: [video link]. There are three key steps.

First, you need the content. In this case, that means writing the questions about people's personal view of abortion and God's view. Both questions should go into a single "block." You have seen how to create blocks and organize content within Qualtrics during previous chapters.

Second, you need to set up the experimental conditions. That means copying the block you created and changing the order of the questions so there are two paths through the survey: one where participants see the question about their own views first and another where they see the question about God's views first. Once you have copied the block, changing the order is as simple as dragging and dropping the items.

Third, you need to implement random assignment—the essential feature for making this an experiment rather than a correlational study. Random assignment in Qualtrics relies on a tool called the "Randomizer" within the Survey Flow. After you add the Randomizer, you can drag your two blocks into the randomization. Then, toggle the settings so the Randomizer randomly selects just one condition for each participant (Figure 7.3).

With these three simple steps, you have created the core of an experiment. To this core, you can add a welcome message, some demographic questions, and an end of study message. Then you are ready to gather data. You could also easily tweak this project to study other questions you are curious about.

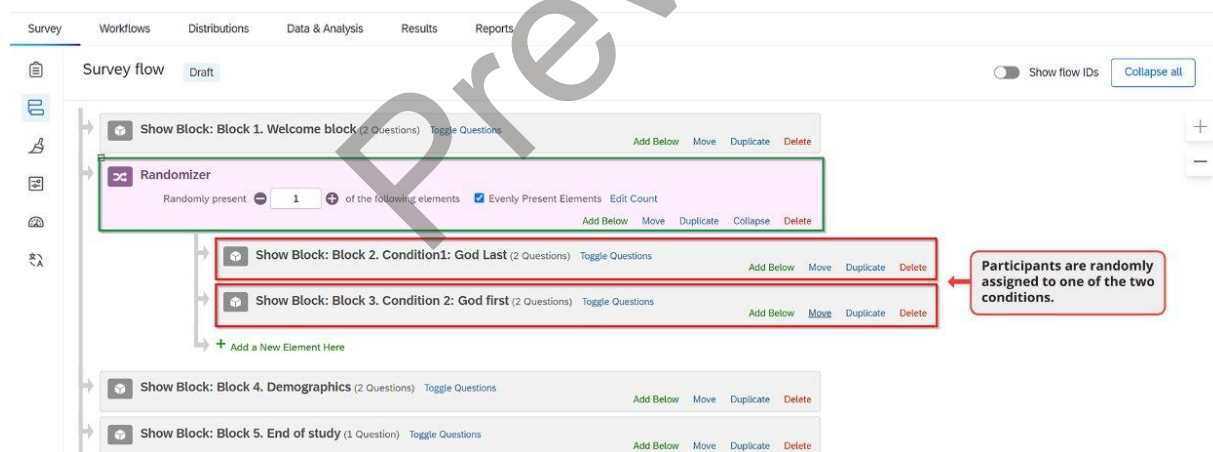


Figure 7.3 Randomization of participants to blocks within the Qualtrics ‘Survey flow.’

Creating Variations on the God Views Experiment

The God Views experiment works by activating a specific mental framework before asking people about their opinion. When people consider God's perspective on abortion before their

own view, they express less support for abortion than when they answer in the reverse order. The activation of this religious or moral framework—what psychologists call priming—influences how people subsequently evaluate the issue.

To create your own version of this experiment, you could keep the structure but change the dependent variable. For example, you could ask people about God's views on climate change before measuring their environmental attitudes. You could inquire about God's views on wealth distribution or taxation before asking people about their perspectives on those issues. Or, you could ask about God's views on capital punishment before asking people about the death penalty. The point is: the experimental structure remains the same, but the topic changes.

As an alternative to changing the dependent variable, you could modify the independent variable. Instead of activating people's mental framework for religion, you might ask them to reflect on what their parents would think about an issue before stating their own opinion. You could also ask people to consider how future generations or their political party would evaluate an issue before giving their judgment. Although the group or entity people are asked to consider may change, the manipulations follow the same principle: change the mental context in which people form and express their attitudes. By systematically varying the context people are asked to consider, you can understand how different social contexts shape people's judgments.

In the next module, we will explore another variation of experimental design by studying how perspective-taking affects moral judgments in the Heinz dilemma.



Programming the Gods View study as a Qualtrics experiment

Once you have created your Qualtrics version of the Gods View study, create an anonymous link and paste it in your portfolio. Do the same for the second study in which you change either the independent or dependent variable. Then write a few sentences describing what your experiment investigates and what you expect to find.

Guided Project: Can Perspective-Taking Shift Moral Judgment?

Module 7.2

Conduct an experimental study on perspective-taking and moral judgment

In this chapter, you have learned how experiments establish cause-and-effect relationships. Now you can put your understanding to work in a guided project.

For this project, we will return to the Heinz dilemma. However, instead of describing people's responses (as you did in Chapter 3) or exploring which moral foundations correlate with people's judgments (as you did in Chapter 5), we are going to test whether people's judgments change when they engage in perspective-taking. This project will give you hands-on experience with the important elements of experimental research: implementing an experimental manipulation, randomly assigning participants to conditions, analyzing group differences, and interpreting causal effects.

Each step of the process is explained in the accompanying video for this project: [\[link to video\]](#).

Project Goals

The primary goal of this project is to introduce you to the basics of experimental design. By working through a guided project, you will learn to form your own experimental hypotheses, how to program a between-subjects experiment in Qualtrics, how to analyze a two-groups design, and how to create figures to visualize the results. Let's begin with the background of this study.

Part 1: Understanding Perspective-Taking

In Chapter 3, you conducted a descriptive study examining how people respond to the Heinz dilemma. You may find it helpful to review that study before beginning this project.

As a reminder, the findings in Chapter 3 revealed that when people were asked whether Heinz should steal the drug or not, they were evenly split. Slightly more than half of people said Heinz should not have stolen the drug. Despite this tendency, people were also sympathetic to Heinz's situation, rating his behavior as somewhat morally acceptable. In this experiment, you will explore what might cause some people to support Heinz's actions and others to oppose them.

One factor that influences people's judgments across a range of moral situations is how much they empathize with others. Just as you might better understand a friend's decision after "putting yourself in their shoes," asking people to take Heinz's perspective might lead them to view his actions more sympathetically. Indeed, perspective-taking is a powerful psychological tool that has been shown to increase empathy and compassion for others, reduce prejudice, lead people from different groups to see divisive issues more similarly, promote helping behavior, and increase recognition of others' hardships (Batson et al., 1997; Batson et al., 2005; Davis et al., 1996; Galinsky & Moskowitz, 2000; Simon et al., 2019; Todd et al., 2012).

The basic idea is that when people imagine another person's perspective, it's easier to consider that person's feelings and circumstances. This often leads to more empathetic judgments of their behavior. We will draw on this well-established area of research to test if perspective-taking affects people's judgments of Heinz.

Before developing your hypotheses, explore some background research on perspective-taking. Take 15 minutes or so to read about perspective-taking and empathy on Google Scholar, looking at articles by researchers like Batson, Davis, or Galinsky (all cited above). Consider how perspective-taking might influence moral judgments in the Heinz dilemma.

After familiarizing yourself with this research, develop your hypotheses about how perspective-taking might affect people's responses to the Heinz dilemma. Consider these questions:

1. How do you predict perspective-taking will affect whether people think Heinz should steal the drug (the yes/no decision)?
2. How do you think perspective-taking will affect ratings of how morally acceptable it is for Heinz to steal the drug (the 1-7 scale)?

For each prediction, write a few sentences in your portfolio explaining your reasoning. Consider which aspects of Heinz's situation might become more salient when people take his perspective. How might understanding his emotional state affect moral judgments? Remember, in experimental research, researchers make causal predictions (e.g., taking Heinz's perspective should cause people to judge his actions differently than remaining objective).

Part 2: Research Design, Materials, and Methods

Now that you have hypotheses, it is time to test them. Here is an overview of the study materials and experimental design.

As with all other experiments, the key is to create a different experience for participants across conditions. In this study, there are two. In the Perspective-Taking condition, people are presented with the following instructions before they read the Heinz dilemma: "On the next page, you will be presented with a brief scenario. Please read the scenario carefully. As you are reading, try to visualize clearly and vividly what the main character, Heinz, is thinking, feeling, and experiencing. Look at the world through his eyes and walk in his shoes."

In the Objective condition, people are presented with the similar but slightly different instructions that follow, before they read about Heinz: "On the next page, you will be presented with a brief scenario. Please read the scenario carefully. As you are reading the scenario, try to remain objective and emotionally detached. Try not to get caught up in what the main character, Heinz, might be thinking, feeling, and experiencing."

After reading these instructions, participants in both conditions are presented with the Heinz dilemma and asked to write for 2-3 minutes, following instructions from their assigned condition. Finally, the participants answer questions about Heinz's actions. Figure 7.4 presents an overview of the design.

Dependent Measures

The dependent variables in this study are the same items used in previous chapters to measure people's reactions to the Heinz dilemma. First, "Should Heinz have stolen the drug for his wife?" (Yes/No). Second, "How morally acceptable was it for Heinz to steal the drug?" (1 = Not at all acceptable to 7 = Completely acceptable).

Perspective-Taking Experiment Design

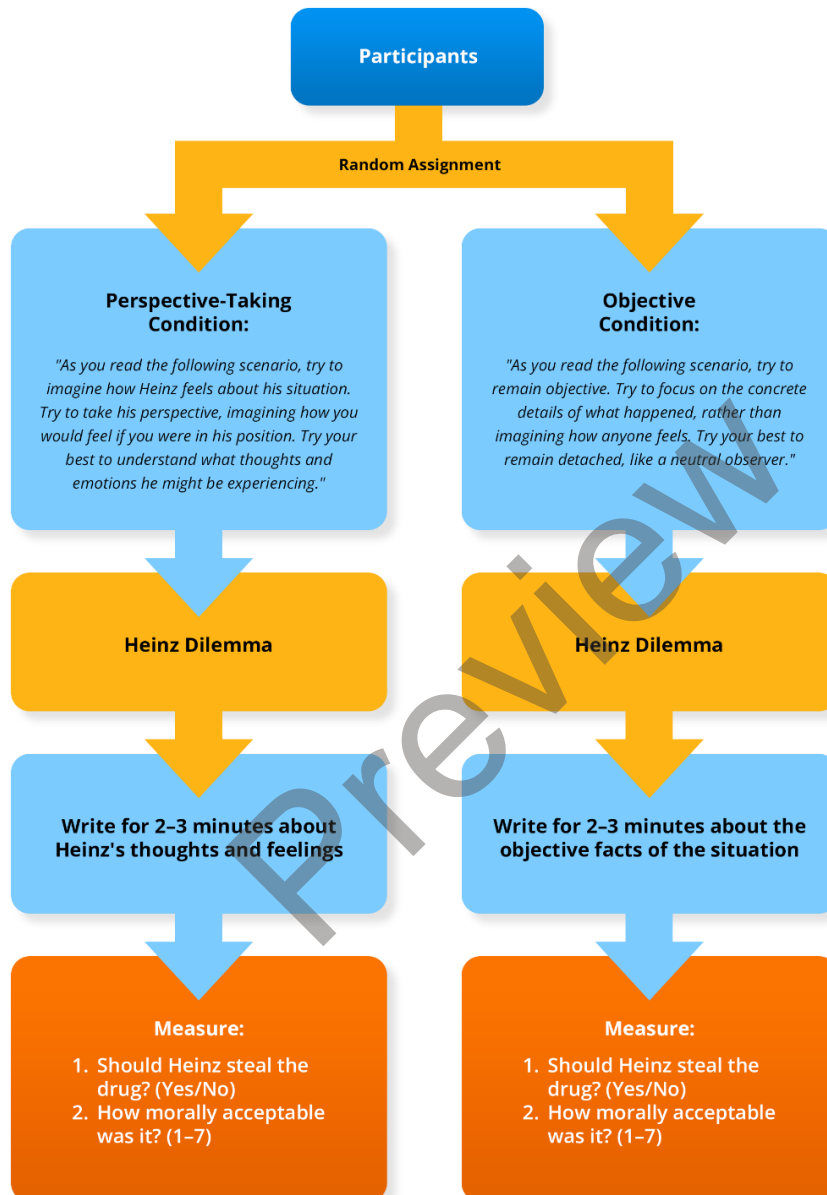


Figure 7.4. The design of the perspective-taking experiment.

Creating the study

To examine how this experiment was implemented, download the Qualtrics survey file from the OSF project page: osf.io/a8kev. The file is named “RITC_SURVEY_CH07_HeinzPerspectiveTaking.qsf” and it is within the “Ch. 7- Experimental Research” folder. Upload it to your Qualtrics account.

Once you have the survey open, review the questionnaires and materials. Examine how the manipulation and measures are structured. Then look at the survey flow. The instructional video highlights the key features of the design.

One indispensable part of the experiment is random assignment to conditions. When participants start the study, Qualtrics assigns them to either the perspective-taking or objective condition with equal probability (like flipping a coin). You can see the random assignment within the survey flow. Notice how each participant is randomly assigned to one of the two conditions (Figure 7.5).

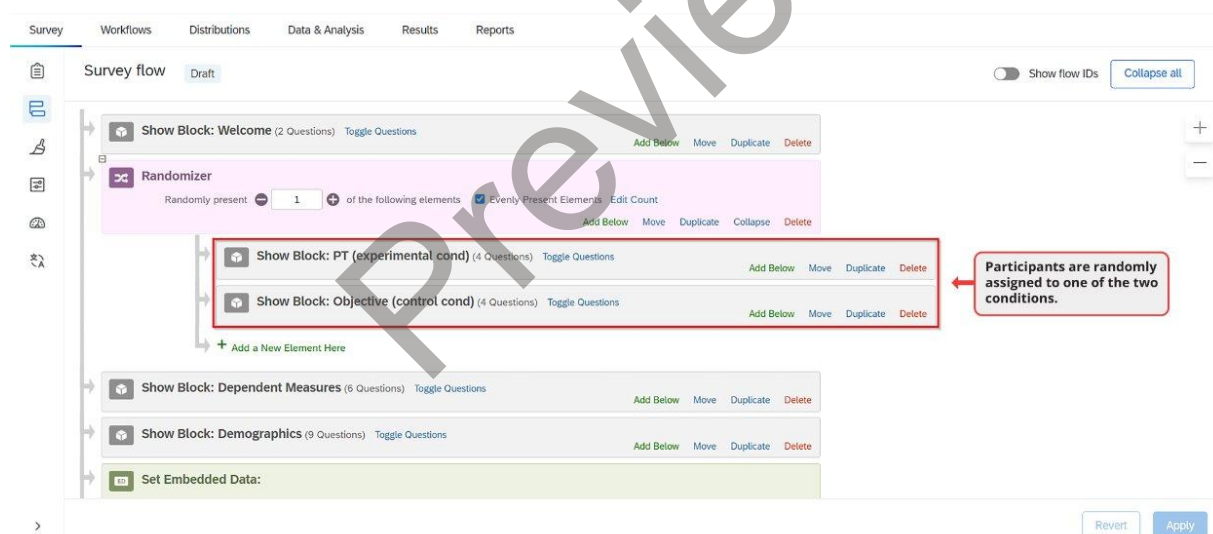


Figure 7.5. Randomization within the perspective-taking experiment.

Part 3: Data Collection

After designing the study, the next step is to gather data. As with all guided projects, we have done that for you.

We gathered data from 100 participants on Connect. We paid each person \$0.75 for a ~5-minute study. To analyze the data, download the “RITC_DATA_CH07_HeinzBtwnSubjects.sav” file from the OSF page. Like the survey file you downloaded, the data is in the folder labeled "Ch 7 – Experimental Research."

Part 4: Analyzing What You Found

This experiment has two outcome measures that require different statistical tests. For the yes/no decision about whether Heinz should steal the drug, you need to conduct a chi-square test. This test will compare the proportion of participants who said "yes" in each condition. The video for this activity will show you how to conduct the analysis and create a figure showing the percentage of "yes" responses by condition. You can also follow the instructions in HOW TO Box 7.1.

For the moral acceptability ratings (measured on a 1-7 scale), you need to conduct an independent samples *t*-test. This test compares the average ratings between the perspective-taking and objective conditions. Once again, you can follow the video or use HOW TO Box 7.2 for instructions. You should conduct the test and create a figure for the results.

After running these analyses, you will interpret whether perspective-taking had a significant effect on either outcome measure. A significant result ($p < .05$) suggests the manipulation caused changes in participants' judgments. The direction of any differences tells you whether perspective-taking made participants more or less supportive of Heinz's actions. Pay attention to whether the results align with your hypotheses—did perspective-taking influence moral judgments in the way you predicted?

HOW TO: Analyze Yes/No Responses with Chi-Square Test

Open the dataset

- Open SPSS and load the "RITC_DATA_CH07_HeinzBtwnSubjects.sav" file
- Check that the "Steal" variable is coded (1 = Yes, 0 = No)
- Verify that "Condition" is coded (1 = Perspective-Taking, 0 = Control/Objective)

Run the Chi-Square analysis

- Click on "Analyze" in the top menu
- Select "Descriptive Statistics > Crosstabs"
- Move "Steal" to the "Rows" box
- Move "Condition" to the "Columns" box
- Click the "Statistics" button
 - Check "Chi-square"
 - Click "Continue"
- Click the "Cells" button
 - Check "Column percentages"
 - Click "Continue"
- Click "OK" to run the analysis

Create a bar chart showing the percentage of "Yes" responses

- Click on "Graphs" in the top menu --> select "Chart Builder"
- Choose "Bar" chart from the gallery
- Drag a Simple Bar chart to the canvas
- Drag "Condition" to the X-Axis
- Drag "Steal?" to the Y-Axis
- Click "OK" to create the chart

Interpret the results

- In the Chi-Square Tests table, find the Pearson Chi-Square value
- Look at the "Asymptotic Significance (2-sided)" value (p-value)
- If $p < .05$, there is a significant difference between conditions
- In the Crosstab table, examine the column percentages
 - Compare the percentage of "Yes" responses in each condition
- Determine if perspective-taking increased or decreased support for stealing
- Compare the findings to your original hypothesis

Box 7.1. Instructions for the chi-square in the perspective taking experiment.

HOW TO: Analyze Moral Acceptability Ratings with a t-test

Open the dataset

- Open the "RITC_DATA_CH07_HeinzBtwnSubjects.sav" file if it is not already open

Run the Independent Samples t-Test

- Click on "Analyze" in the top menu
- Select "Compare Means > Independent-Samples T Test"
- Move "Acceptability" to the "Test Variable(s)" box
- Move "Condition" to the "Grouping Variable" box
- Click the "Define Groups" button
 - Enter "1" for Group 1 (Perspective-Taking)
 - Enter "0" for Group 2 (Objective)
 - Click "Continue"
- Click "OK" to run the analysis

Create a bar chart with error bars

- Click on "Graphs" in the top menu --> Select "Chart Builder"
- Choose "Bar" chart from the gallery
- Drag a Simple Bar chart to the canvas
- Drag "Condition" to the X-Axis
- Drag "Acceptability" to the Y-Axis
- Click the "Element Properties" button
 - Click on the "Error Bars" tab
 - Select "Standard Error" with Multiplier = 1
- Click "OK" to create the chart
- Double-click the chart to edit in Chart Editor
 - Add appropriate titles and labels
 - Click "Close" when finished

Interpret the results

- In the "Independent Samples Test" table, locate the t-value, df, and Sig. (2-tailed)
- Look at the "Sig. (2-tailed)" value (p-value) for the t-test
 - If $p < .05$, there is a significant difference between conditions
- Examine your bar chart to see the direction of the effect
 - Which condition shows higher moral acceptability ratings?
- Determine if perspective-taking increased or decreased moral acceptability ratings
- Compare your findings to your original hypothesis

Box 7.2. Instructions for conducting a t-test in the perspective taking experiment.



Portfolio Entry: Report a Two Group Randomized Experiment

Once you have conducted the analyses, paste the output and the graphs you created into your portfolio. Using the templates for reporting t -tests and chi-square results you have seen in previous chapters, report and interpret the results of the experiment. Did the results align with your hypotheses?

Your Turn: Exploring Perspective-Taking Effects

Now that you have seen how perspective-taking can influence moral judgments, you are ready to investigate how this manipulation might affect other judgments and behaviors. The manipulation from this experiment—asking people to imagine someone else's thoughts and feelings versus asking them to remain objective—can be applied to many different scenarios.

For instance, you might examine how perspective-taking affects judgments in other ethical dilemmas. What happens when people take the perspective of someone who cheated on an exam because they needed a passing grade to keep their scholarship? Or does perspective-taking make people more sympathetic toward someone who lied to protect a friend? The same manipulation that influenced judgments about Heinz might shape how people view other moral transgressions.

Perspective-taking might also influence how people judge controversial policies. You could present participants with a story about someone affected by immigration policies, healthcare costs, or educational inequalities. Would taking that person's perspective change support for the associated policy? Answering these questions could shed light on how personal stories affect people's attitudes.

Another possibility is to investigate perspective-taking in interpersonal conflicts. You could describe a disagreement between roommates, coworkers, or romantic partners. Maybe you have a story from your own life. If so, you can explore whether taking one person's perspective changes how people assign blame in the situation.

Whatever topic you choose, you can use the same basic experimental structure. Some participants receive instructions to take the perspective of the person in your scenario, while others are told to remain objective. Then, you present participants with the scenario and measure some dependent variable that perspective-taking might plausibly affect.

You can use the Qualtrics survey from the Heinz experiment as a template. Simply replace the Heinz dilemma with your chosen scenario and modify the dependent measures to fit your research question. The video that accompanies this activity shows you how to adapt the materials while maintaining the essential structure of a two-group experiment.



Portfolio Entry: Design Your Own Two-Groups Between-Subjects Experiment

Create a Qualtrics perspective taking experiment. It can relate to any question of interest where perspective taking may cause a change in judgment. Paste a link to the experiment in your portfolio.

Variations on Experimental Design: Repeated Measures Experiments

Module 7.3

Explore repeated measures designs by conducting a guided project on how different consequences influence moral decision-making.

We have discussed experiments where different groups of participants experience different conditions—what researchers call a between-subjects design. In the God Views study, for instance, each participant either answered questions about God's views first or their own views first. In clinical trials, each patient receives either the drug or the placebo, but not both. There is, however, another way to structure experiments—one where each participant experiences every experimental condition in the study.

To demonstrate how this works, consider a study examining how background music affects cognitive performance. In a **between-subjects** design, some participants are randomly assigned to listen to classical music while solving math problems and others to solve problems in silence. Although this design is effective, it requires a lot of participants. If the sample size is small, random assignment may not effectively distribute individual differences across conditions. Remember, with only a few participants in each condition, it is possible to get “unlucky” and end up with groups that differ in some important way, even after random assignment.

This is where **within-subjects** designs (also called **repeated measures**) offer an advantage. Instead of comparing different groups, researchers can have each person solve problems with music *and* in silence. Within this design, it is possible to see how music affects each person's performance relative to their own baseline ability. In other words, each person serves as their own control. This makes it possible to see how the same person performs under different conditions, rather than comparing different groups of people (Figure 7.6). As a result, fewer participants are needed.



Figure 7.6. Differences between independent groups (between-subjects) and repeated measures (within-subjects) experiments.

However, within-subjects designs face their own challenges. And the main challenge is **order effects**—the possibility that experiencing one condition affects how people respond to later conditions. In the math study, for instance, you might expect people to get better at math problems due to practice and regardless of the music condition they are assigned to. Or, you might see participants get tired and perform worse later in the study than at the start. To address these concerns, repeated measures experiments need to **counterbalance** the order of conditions across participants. This means some people will be assigned music first and silence second,

while others get the opposite order. Counterbalancing ensures that the order of the conditions cannot explain the study's outcomes.

Repeated measures are not suited to all situations. For instance, consider Stanley Milgram's famous obedience experiments, where participants were instructed to deliver increasingly strong electric shocks to a stranger (actually an actor who received no real shocks). This study cannot work as a repeated measures design. Once a participant has experienced the study, they cannot forget the situation was staged or that the shocks were fake. Their experience in the first condition permanently alters how they respond in subsequent conditions.

Similarly, most studies that involve deception, require a between-subjects design. This is because the knowledge people gain in one condition often creates a lasting change that would contaminate how they respond to subsequent conditions. Whenever this is the case, behavioral scientists must rely on between subjects designs, despite their need for larger sample sizes.

Guided Research Project: How Consequences Shape Moral Judgments

In previous studies with the Heinz dilemma, stealing the drug always saved Heinz's wife. But in the real world, outcomes are often less certain. Sometimes, a novel treatment only eases people's pain rather than providing a cure. At other times, experimental treatments offer uncertain benefits. How do these different outcomes affect whether people think stealing is justified?

In this guided research project, you will gain experience with the key elements of a repeated measures design: creating multiple versions of a scenario, counterbalancing their presentation, and analyzing how the same people respond across conditions. The video that accompanies this project will guide you through each step: [video link].

Project Goals

This project has several goals.

First, you will generate hypotheses about people's judgments under different conditions in the Heinz dilemma. This will further your ability to think about research questions and develop testable predictions.

Second, you will program a repeated measures experiment with three conditions. Programming the experiment will give you further knowledge of how to use Qualtrics and hands-on experience with the practice of counterbalancing measures.

Third, this project will extend your knowledge of statistical analyses. You will learn how to analyze within-subjects data and report the results.

Finally, as in other guided assignments, you will practice creating a figure to visualize the results. Overall, this project will expand your knowledge of experimental design while allowing you to practice some of the basic skills required for experimental research.

Part 1: Understanding the Repeated Measures Design

In this study, participants were presented with three versions of the Heinz dilemma that differed only in the consequences of the drug Heinz considered stealing. In one version, the drug would save his wife. In another version, the drug was described as easing pain and suffering. Finally, in the third version, the drug was experimental with uncertain benefits.

Because this was a repeated measures design, *each participant saw all three versions of the experiment*. This design allowed us to examine how the same person's moral judgments changed as the consequences of stealing varied.

Developing Your Hypotheses

How do you think people will react to the different versions of the scenario? Develop a few hypotheses about whether you expect differences between conditions and which conditions you think will lead to the most favorable ratings of Heinz's behavior.

Write a paragraph in your portfolio explaining your predictions and the thinking behind them. Remember, participants see all conditions of the experiment, so you are predicting how the same person's judgments will change across different versions of the scenario.

Part 2: Research Design, Materials, and Methods

The key to this experiment is creating three versions of the dilemma that differ only in the consequences of stealing the drug. The basic scenario remains identical across versions—Heinz

cannot afford the drug, the company will not lower the price, he considers stealing it. The only thing that varies is what the drug will do for his wife. Across conditions participants were told to imagine:

Life-Saving Version: "As a result of Heinz's behavior, his wife is cured and lives a long life."

Pain-Relief Version: "As a result of Heinz's behavior, his wife spends her last few weeks free of pain but the drug does not save her life."

Experimental Version: "As a result of Heinz's behavior, his wife will be able to take the experimental drug but its benefits are uncertain."

After reading each version, participants answered two questions: 1) Should Heinz steal the drug? (Yes/No), and 2) How morally acceptable would it be for Heinz to steal the drug? (1 = Not at all acceptable to 7 = Completely acceptable). These were the same dependent measures used in previous experiments.

Counterbalancing

Because this is a repeated measures design, each participant saw all three versions of the scenario. However, the order of the scenarios was randomly assigned to control for order effects.

Someone who saw the life-saving version first might judge subsequent versions more harshly by comparison. Or participants might become more accepting of stealing as they think repeatedly about Heinz's situation, regardless of the consequences. To avoid these and all other "order effects" we used Qualtrics's randomization tool to present the three versions in random order (Figure 7.7).

The randomization tool ensured that each version of the dilemma appeared equally as often in the first, second, and third position across all participants. The instructional video for this project shows you how to implement counterbalancing in Qualtrics.

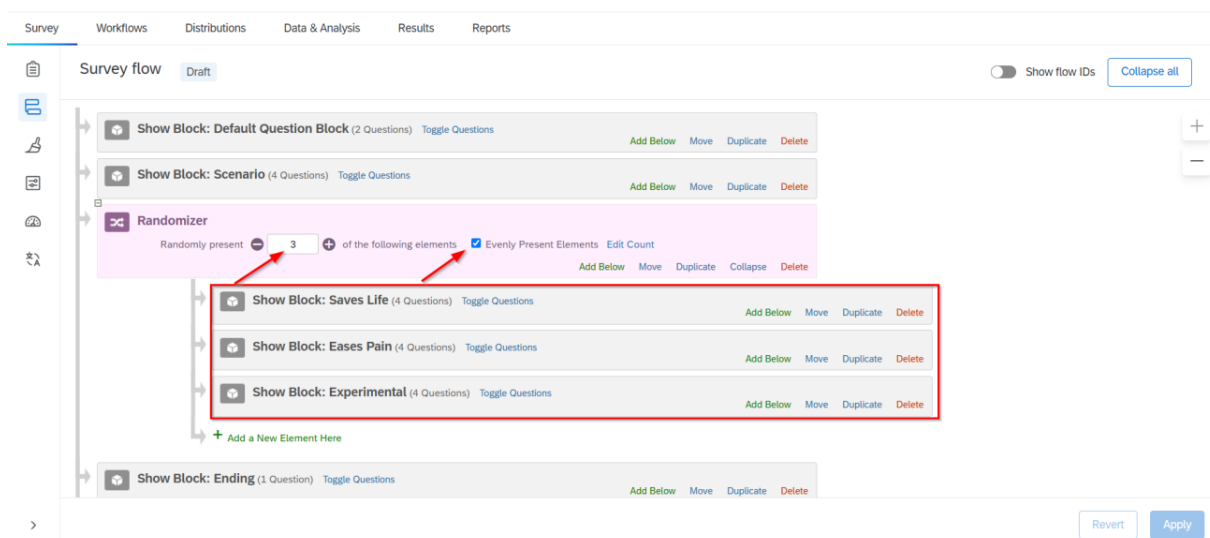


Figure 7.7. The three conditions are presented in a randomly determined order for each participant.

Creating the Study

To see how this experiment was implemented in Qualtrics download the “RITC_SURVEY_CH07_HeinzWithinSubjects.qsf” file from the OSF page. It is in the folder labeled “Ch. 7 – Experimental Research.”

Once you upload this file into Qualtrics, look around. Notice how all participants read one version of the dilemma and then were asked to imagine one of three different outcomes before they answered the dependent variables (Figure 7.6). Navigate to the survey flow and examine the randomizer. Notice how all three elements are presented evenly across participants. Once you are done reviewing the survey, it is time to get the data.

Part 3: Data Collection

We gathered data from 50 participants on Connect—half the number compared to the between-subjects design. We paid each person \$0.50 for a ~3-minute study.

While the data collection itself was simple, preparing the file for analysis required one important step not necessary in any study you have conducted until now: we had to ask Qualtrics for the order each participant viewed the manipulations in. Adding this information to the data file required just a few extra clicks while downloading the data, but without it we would not have

been able to create a condition variable that identifies the order each participant received when working through the study. The video that accompanies this assignment explains how to add the viewing order to a data file.

When you are ready to work with the data, find the folder labeled "Ch 7 – Experimental Research" on the OSF project page. Then, download the “RITC_DATA_CH07_WithinSubjects.sav” file.

Part 4: Analyzing What You Found

Within-subjects experiments require a different analytical approach than between-subjects designs because the same people respond across different conditions.

To analyze the data from a within-subjects experiment, it’s common to use a *repeated measures analysis of variance* (ANOVA). The repeated measures ANOVA compares participants’ responses within different conditions to see if there are significant differences. HOW TO Box 7.3 provides step by step instructions for conducting this analysis, as does the instructional video online.

When examining the statistical output, you need to know what to look for and why.

First, examine the table labeled "Tests of Within-Subjects Effects." Find the row with your factor name, in this case "Consequences", and check the significance value. You should see a p-value below .05. If $p < .05$, there is a significant overall effect, meaning that at least one condition differs from the others. Later, you will use the F-statistic and degrees of freedom (df) under the row of “Sphericity assumed” for reporting the results.

To understand the differences between conditions, review the descriptive statistics table. This table provides the average response for each condition, allowing you to see the pattern of responses. To see which means are significantly different from one another, analyze the **Pairwise Comparisons Table** (Figure 7.8). This table shows which specific conditions differ from each other with a $p < .05$ indicating statistical significance. Note which conditions differ and which do not.

HOW TO: Analyze Within-Subjects Experiment Data

Open the dataset

- Open SPSS and load the "RITC_DATA_CH07_HeinzWithinSubjects.sav" file

Run the Repeated Measures ANOVA

- Click on "Analyze" in the top menu
- Select "General Linear Model" > "Repeated Measures..."
- In the Define Factor dialog box:
 - Type "Consequences" as the factor name
 - Enter "3" for Number of Levels
 - Click "Add" then "Define"
- Move the three acceptable variables (e.g., acceptable_SL) to the Within-Subjects Variables box

Request Estimated Marginal Means

- Click "Options" to open the Options dialog
- In the "Estimated Marginal Means" section:
 - Select your factor "Consequences" and click "Add"
 - Check "Compare main effects"
 - Select "Bonferroni" from the dropdown for multiple comparisons
- Also check "Descriptive statistics" to get means and SDs
- Click "Continue" and then "OK" to run the analysis

Interpret the results

- First check the "Tests of Within-Subjects Effects" table for overall significance
- Then examine the "Pairwise Comparisons" table to see which specific conditions differ
- Use the mean differences and p-values to determine statistical significance

Box 7.3. Instructions for conducting a repeated measures analysis of variance.

Pairwise Comparisons						
Measure: MEASURE_1						
(I) consequences	(J) consequences	Mean Difference (I-J)	Std. Error	Sig. ^b	95% Confidence Interval for Difference ^b	
					Lower Bound	Upper Bound
1	2	.585*	.165	.002	.177	.993
	3	.377*	.133	.019	.049	.705
2	1	-.585*	.165	.002	-.993	-.177
	3	-.208	.156	.564	-.592	.177
3	1	-.377*	.133	.019	-.705	-.049
	2	.208	.156	.564	-.177	.592

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

Figure 7.8. The table of pairwise comparisons shows which condition means are significantly different from one another.

Creating a Figure

Your results should include a figure showing how moral judgments changed across the three conditions. HOW TO Box 7.4 provides instructions for how to create this figure. Include error bars and make sure axes are clearly labeled. Your final result should look similar to Figure 7.9.

HOW TO: Create a Figure for Within-Subjects Results

Create a Bar Chart

- Click on "Graphs" in the top menu
- In the dropdown select the "Bar..." Option
- In the pop up, select "Simple" and under "Data in Chart Are" choose "Summaries of separate variables"
- Select "Define"

Set Up the Variables

- Drag all three variables to the "Bars Represent" box
- Make sure the variables represent MEANS

Add Error Bars

- Click on the "Options" button
- Select the "Display Error Bars" box
- Choose "Standard Error" and set the multiplier to "1"
- Click "Continue" then "OK" to add the error bars

Add Labels

- Double click on the Chart
- In the pop up window, right click --> Show Data Labels

Box 7.4. Instructions for how to create a bar chart for the results of the repeated measures experiment.

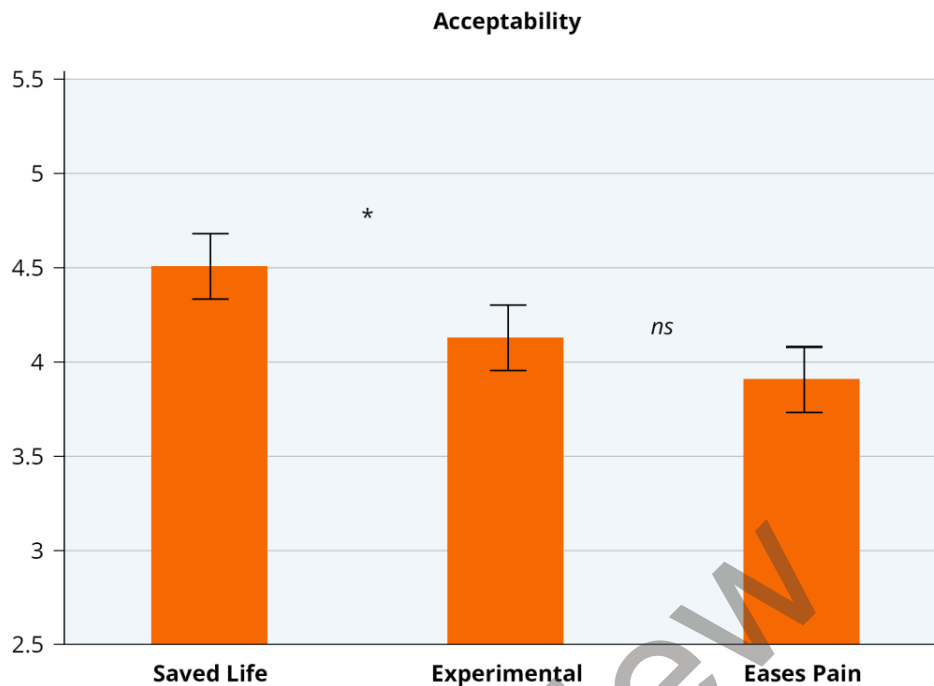


Figure 7.9. Results of the within subjects experimental design.

Results and Interpretation

After analyzing the data, you need to interpret what your findings mean and write them up. You will also compare your results to your original hypotheses.

When writing up within-subjects results, researchers typically report both the overall ANOVA and the specific contrasts between conditions. Here's an example:

"A repeated measures ANOVA revealed that the drug's consequences significantly affected people's moral judgments, $F(2, 104) = 7.65, p < .05$.

A post-hoc analysis with Bonferroni corrections indicated that participants found stealing more acceptable when the drug saved Heinz's wife ($M = 4.57, SD = 2.30$) compared to when it eased her pain ($M = 3.98, SD = 2.28$), $p < .01$. They also judged stealing more acceptable when the drug was life-saving than when it was experimental ($M = 4.19, SD = 2.21$), $p = .02$. The difference between a drug that eased pain and was experimental was not significant, $p > .56$."

Notice how this write-up includes both the overall test and the specific comparisons between conditions. The overall test information comes from the “Tests of Within Subjects Effects” table. Each conditions’ mean and standard deviation is taken from the descriptive statistics table, and the p values come from the Pairwise Comparisons table. With these statistics, you have the information necessary to make sense of the study.



**Portfolio Entry: Reporting the Results of a Repeated-Measures Experiments:
Heinz Dilemma and Different Drug Outcomes**

Describe your original hypotheses. Then, report the results of the repeated measures ANOVA. Your results section should include the figure showing differences in condition means. After reporting the statistical results, explain what they mean in plain language. Are people more willing to justify breaking the law when the benefits are more certain? Consider how your results connect to broader questions about how people make moral decisions.

Beyond Simple Experiments: Factorial Designs

Module 7.4

Investigate interactions by running a factorial experiment on perspective-taking and moral judgments.

Every experiment you have seen so far manipulated just one variable. Yet, in the real world, multiple factors often work together to influence people's thoughts, feelings, and behaviors. Factorial designs allow researchers to examine this kind of complexity. Instead of asking "Does X cause Y?" factorial designs ask questions like "Under what conditions does X cause Y?" or "For whom is the effect of X on Y strongest?"

Think about the finding from earlier showing that people express less support for abortion after considering God's perspective first. This effect probably is not the same for everyone. Someone who does not believe in God might be unaffected by the manipulation. But for a deeply religious person, considering God's perspective might have a big effect on the personal opinion they report.

To investigate this idea, a researcher would need to examine both variables together—thinking about God's views AND how religious people are. This could be as simple as adding a measure of religiosity to the experiment you viewed earlier. After the main task, participants could answer a few questions about their religious beliefs while completing other demographic information. Based on the responses, the researcher could classify each person as either high or low in religiosity.

Figure 7.10 shows what the results of this experiment might look like. The graph reveals several important patterns that illustrate why factorial designs are so important.

The most interesting pattern is how the effect of question order differs between highly religious and less religious participants. For highly religious people, thinking about God's views first reduces support for abortion—notice the difference between the blue and orange bars on the left side of the graph. But for less religious people, thinking about God's views has no effect—the bars on the right side are about the same.

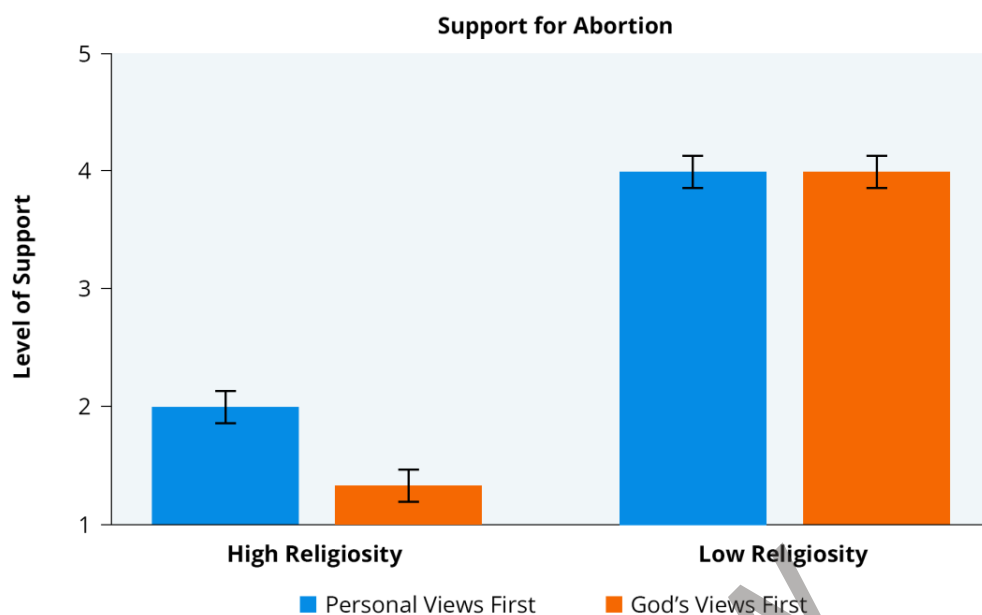


Figure 7.10 Results of a factorial study showing a main effect of religiosity and an interaction between religiosity and the God perspective manipulation.

This pattern is called an **interaction**. An interaction occurs when the effect of one variable (God's View manipulation) *depends on* the **level of another variable** (religiosity). This interaction indicates that thinking about God's views does not affect everyone equally. It decreases religious people's support for abortion but has no effect on less religious people's support. By using a factorial design such as this, researchers can specify for whom an effect occurs and how strongly.

Beyond the interaction, a factorial design also provides information about what are called **main effects**. A main effect is simply the effect one independent variable has on the dependent variable, while ignoring the other independent variable in the design. You can see a main effect of religiosity in Figure 7.10 by noticing the overall difference between the left and right sides of the graph. Regardless of when they think about God's views (first or second), highly religious people express less support for abortion than less religious people. If you averaged the two bars on the left side of the graph, averaged the two bars on the right side, and performed the

appropriate statistical test, you would find a significant difference, indicating a main effect of religiosity.

Guided Research Project: When Does Perspective-Taking Work?

In this guided project, you will examine how multiple factors work together to influence behavior using a factorial design. Once again, you will return to perspective-taking in the Heinz dilemma. But this time, you will investigate whether its effects depend on Heinz's economic circumstances. Does perspective-taking make people more sympathetic to Heinz's actions regardless of his wealth, or does his socio-economic status change how people respond to his situation?

This project will give you hands-on experience with the key elements of factorial design: manipulating multiple variables simultaneously and analyzing both main effects and interactions. As with previous projects, the accompanying video provides step-by-step instructions: [video link].

Project Goals

This project has several objectives.

First, you will form hypotheses regarding the interaction between two independent variables and their effects on participants' responses. Second, you will design and implement a 2×2 factorial experiment using Qualtrics. Third, this project will broaden your statistical expertise; you will learn how to conduct and interpret analyses appropriate for factorial designs, including assessing main effects and interactions between variables. Fourth, you will create a visualization that effectively communicates both main effects and interaction effects in your data. Overall, this project will expand your experimental design toolkit while reinforcing fundamental research skills, particularly in the context of multi-factor studies that allow for examining how variables may interact to influence outcomes.

Part 1: Understanding the Factorial Research Question

In the previous study, you saw that taking Heinz's perspective makes people more sympathetic to his actions. But does this effect work equally well in all situations? What if Heinz was relatively wealthy? Would taking his perspective still make people more understanding of his actions?

A factorial design can answer these questions. Instead of just manipulating perspective-taking, it is possible to vary Heinz's economic circumstances. By examining both perspective-taking and wealth simultaneously, we can ask several interesting questions: 1) Does perspective-taking increase support for Heinz's actions regardless of his wealth?, 2) Are people generally more accepting of stealing when someone is poor versus wealthy?, and 3) Does the effect of perspective-taking depend on Heinz's economic circumstances?

This last question addresses the interaction. Maybe perspective-taking is especially powerful when someone is poor but less effective when they have some wealth. Or perhaps perspective-taking helps people understand Heinz's choice regardless of his financial situation.

To get started, generate some hypotheses about the main effects and interaction in this study and enter them in your portfolio.

Part 2: Research Design, Materials, and Methods

After you have developed hypotheses about how perspective-taking and wealth might influence moral judgments, let's examine how to test these predictions using a factorial design. The study requires four versions of the Heinz dilemma that combine the two manipulations.

The Experimental Manipulations

The factorial design requires manipulating two independent variables, creating *four* distinct conditions that participants might be assigned to within the experiment (Figure 7.11).

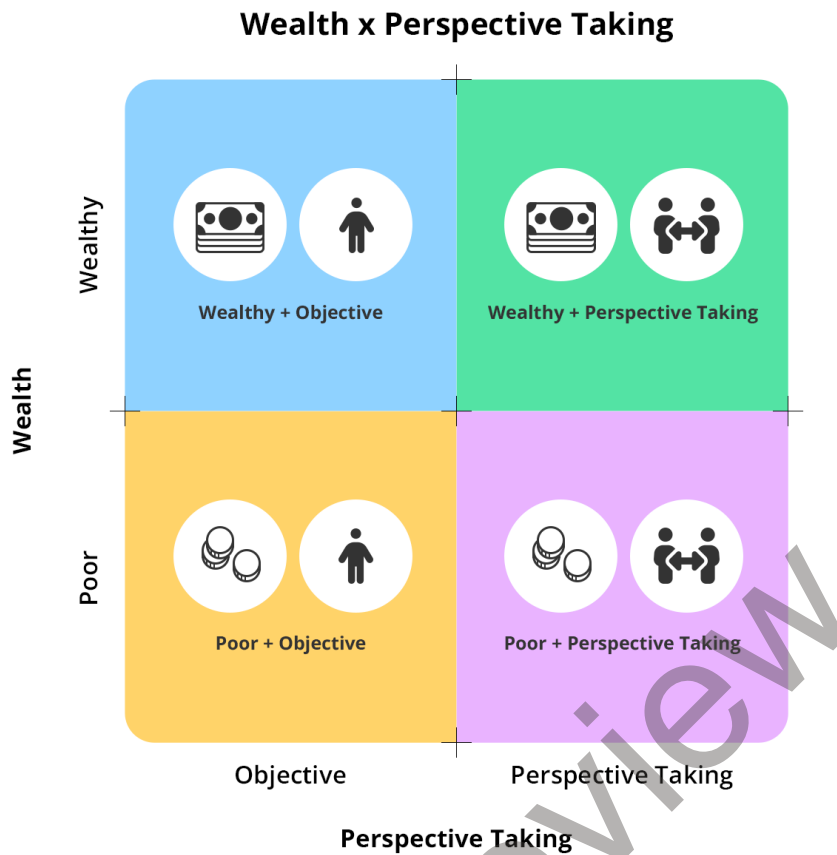


Figure 7.11. A 2×2 factorial design yields four experimental conditions.

The first variable is perspective-taking, which we manipulated through the following instructions.

Perspective-Taking Condition: " On the next page, you will be presented with a brief scenario. Please read the scenario carefully. As you are reading, try to visualize clearly and vividly what the main character, Heinz, is thinking, feeling, and experiencing. Look at the world through his eyes and walk in his shoes."

Objective Condition: " On the next page, you will be presented with a brief scenario. Please read the scenario carefully. As you are reading the scenario, try to remain objective and emotionally detached. Try not to get caught up in what the main character, Heinz, might be thinking, feeling, and experiencing."

The second variable was Heinz's wealth. We manipulated relative wealth within the scenario by telling participants one of two things:

Poor Condition: "Despite working extra jobs and asking everyone they knew for help, Heinz and his wife had only managed to gather about half the money needed for the drug."

Wealthy Condition: "Heinz and his wife had \$200,000 in savings—enough to buy the drug, but using all their money would leave them nothing for her ongoing care and treatment."

Although the wealth condition did not represent truly lavish wealth, we wanted people to think Heinz had enough money that he did not need to steal the drug. When we combined the manipulations, it created the four possible conditions depicted in Figure 7.11

Random Assignment

In a factorial design, random assignment means giving each participant an equal chance of experiencing any of the experimental conditions. After reading their assigned version of the dilemma, participants answered the dependent variables. In this study, we only asked “How morally acceptable would it be for Heinz to steal the drug?” (1 = Not at all acceptable to 7 = Completely acceptable).

Creating the Study

To see the factorial experiment in Qualtrics download the “RITC_SURVEY_CH07_Factorial.qsf” file from the OSF page. It is in the folder labeled “Ch. 7 – Experimental Research.”

Once you upload this file into Qualtrics, notice how we created the four versions of the scenario by combining the two manipulations. Then, examine how random assignment works with multiple variables. Do you notice anything different from past studies? After you review the survey, you are ready for the data.

Part 3: Data Collection

We gathered data from 200 participants on Connect, once again aiming for about 50 people per condition. We paid each person \$0.75 for a ~5-minute study. To analyze the data, download the “RITC_DATA_CH07_HeinzFactorial.sav” file from the OSF page.

As described in the accompanying video, after downloading the data we had to create a variable for each condition. You can find these variables in the data file. You will use them in the analysis.

Part 4: Analyzing What You Found

In factorial experiments, the statistical analysis needs to examine how each variable affects responses independently (main effects) and how they work together (interactions). Both HOW TO Box 7.5 and the instructional video provide instructions for how to conduct the analysis.

HOW TO: Testing Perspective-Taking and Wealth

Generate Research Hypotheses

- Develop hypotheses about perspective-taking's effect when Heinz is wealthy vs. poor
- Consider three types of predictions:
 - Main effect of perspective-taking (improves moral judgments overall)
 - Main effect of wealth (wealthy Heinz judged more harshly)
 - Interaction effect (perspective-taking might help poor Heinz more)
- Write a paragraph explaining your predictions with theoretical rationale

Analyze the Data

- Download the "Heinz Dilemma - Factorial Design.sav" file from OSF
- Run a Two-Way ANOVA in SPSS:
 - Click "Analyze" > "General Linear Model" > "Univariate"
 - Move "Acceptability" to the Dependent Variable box
 - Move both "Perspective" and "Wealth" to the Fixed Factors box
 - In Options, select "Descriptive statistics" and "Estimates of effect size"
 - Select Ok to run the analysis

Box. 7.5. Instructions for analyzing a factorial design.

Analyzing Factorial Effects

The main analysis for factorial designs uses Analysis of Variance (ANOVA), which examines three pieces of information.

First, it tests for a main effect of perspective-taking: whether people judged Heinz's actions differently when taking his perspective versus remaining objective, ignoring both wealth conditions. Second, it tests for a main effect of wealth: whether people judge stealing differently when Heinz is poor versus wealthy, ignoring both perspective-taking conditions. Finally, and most importantly, it tests for an interaction between the variables: whether the effect of perspective-taking differs depending on Heinz's wealth.

The ANOVA provides F-statistics and p-values for each of these effects. A significant result ($p < .05$) for main effects tells you that the variable affected judgments on its own. A significant interaction tells you the variables work together in more complex ways. As shown in Figure 7.12, the main effect of perspective-taking was not significant ($p > .05$). But the main effect of wealth and the interaction both were ($p < .05$). A look at the means for the two wealth conditions shows that people found Heinz's behavior more acceptable when he was poor than relatively wealthy, but this finding is of less interest than the interaction.

Tests of Between-Subjects Effects

Dependent Variable: How morally acceptable was it for Heinz to steal the drug?

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	69.644 ^a	3	23.215	6.308	<.001	.088
Intercept	3625.600	1	3625.600	985.120	<.001	.834
perspective	7.706	1	7.706	2.094	.149	.011
wealth	36.056	1	36.056	9.797	.002	.048
perspective * wealth	24.404	1	24.404	6.631	.011	.033
Error	721.351	196	3.680			
Total	4395.000	200				
Corrected Total	790.995	199				

a. R Squared = .088 (Adjusted R Squared = .074)

Figure 7.12. ANOVA results for a factorial design.

To understand the interaction, you can examine the simple effects tests. These tests tell you whether the effect of one independent variable is significant at a particular level of the other independent variable. For example, is there a difference between perspective-taking and remaining objective when Heinz is described as poor?

In the SPSS output, simple effects tests are labeled “pairwise comparisons.” Examining the comparisons for perspective-taking within each level of wealth reveals that perspective-taking influences people’s judgments of acceptability when Heinz is described as relatively wealthy but not when he is poor (Figure 7.13). What the pattern of means across conditions reveals is that when Heinz is described as relatively wealthy AND people are objective they find his behavior less acceptable than when they engage in perspective-taking.

Estimates

Dependent Variable: How morally acceptable was it for Heinz to steal the drug?

TELLS WHICH PERSPECTIVE CONDITION THE P WAS ASSIGNED TO	TELLS WHICH WEALTH CONDITION THE P WAS ASSIGNED TO	Mean	Std. Error	95% Confidence Interval for	
				Lower Bound	Upper Bound
objective	rich	3.288	.266	2.764	3.813
	poor	4.837	.274	4.296	5.377
perspective-taking	rich	4.380	.271	3.845	4.915
	poor	4.531	.274	3.990	5.071

Pairwise Comparison

Dependent Variable: How morally acceptable was it for Heinz to steal the drugs?

TELLS WHICH WEALTH CONDITION THE P WAS ASSIGNED TO	(I) TELLS WHICH PERSPECTIVE CONDITION THE P WAS ASSIGNED TO	(J) TELLS WHICH PERSPECTIVE CONDITION THE P WAS ASSIGNED TO	Mean Difference (I-J)	Std. Error	Sig. ^b	95% Confidence Interval for Difference ^b	
						Lower Bound	Upper Bound
rich	objective	perspective-taking	-1.092*	.380	.005	-1.841	-.342
	perspective-taking	objective	1.092*	.380	.005	.342	1.841
poor	objective	perspective-taking	.306	.388	.431	-.458	1.070
	perspective-taking	objective	-.306	.388	.431	-1.070	.458

Figure 7.13. Pairwise comparisons test the effect of one independent variable within the level of another.

Creating a Figure

The results of a factorial design are typically displayed in either a line or bar graph. These graphs allow researchers to show how one variable's effect changes across levels of the other variable. For example, the hypothetical results from the God Views study and religion in Figure 7.10 or the results from this study showing how wealth and perspective-taking affect judgments in the Heinz dilemma both show how to visualize an interaction (Figure 7.14).

Using the instructions in HOW TO Box 7.5, create a figure for the Heinz dilemma. Your figure should show moral acceptability ratings on the y-axis (vertical) and wealth on the x-axis (horizontal). There should also be separate lines for the perspective-taking versus objective conditions. The video that accompanies this project also demonstrates how to create these visualizations and interpret patterns of main effects and interactions.

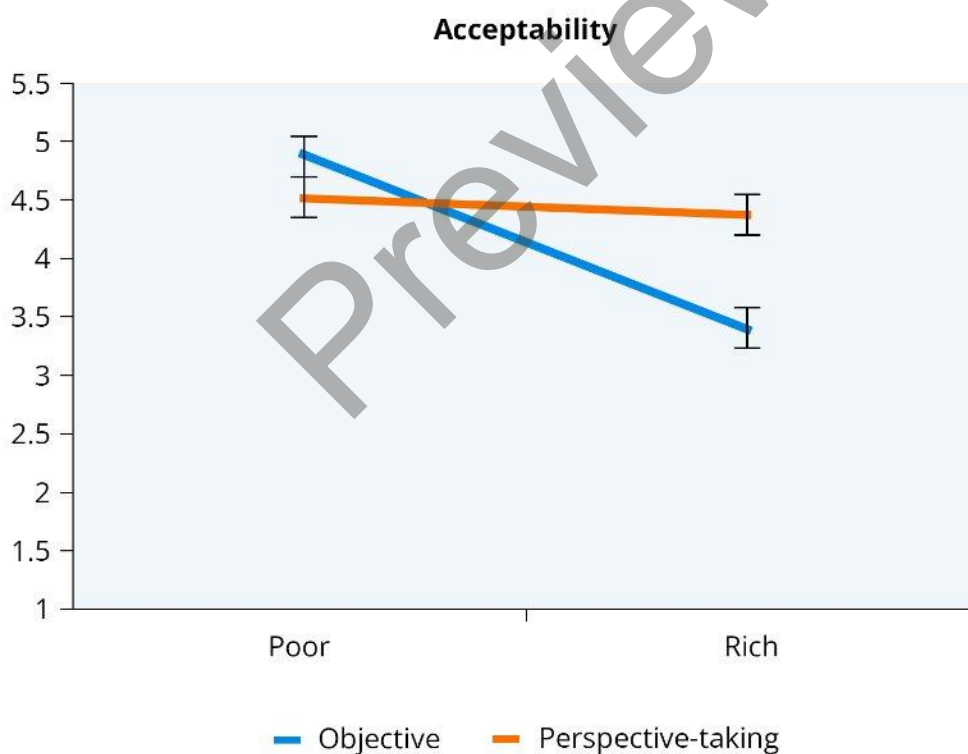


Figure 7.14. The interaction between wealth and perspective-taking in the Heinz dilemma.

Results and Interpretation

Now that you have analyzed the data, interpret what it means. Factorial designs require careful attention to how you explain both main effects and interactions.

Start by comparing your results to your hypotheses. Did perspective-taking affect moral judgments overall? What about Heinz's level of wealth? More importantly, did these factors interact? Did perspective-taking have different effects depending on Heinz's economic circumstances?

When describing factorial results, researchers typically start with the main effects and then move to any interactions. Here's an example:

"A 2×2 ANOVA revealed a significant main effect of wealth, $F(1, 196) = 9.80, p = .002$. People rated stealing as more acceptable when Heinz was poor ($M = 4.68, SD = 1.95$) than when he was wealthy ($M = 3.82, SD = 1.95$). The main effect of perspective-taking was not significant, $F(1, 196) = 2.09, p = .149$. However, there was a significant interaction between wealth and perspective-taking, $F(1, 196) = 6.63, p = .011$."

After reporting that the interaction was significant, it is helpful to break down the effect of one variable at each level of the other. In this study, for instance, wealth affected people's judgments differently depending on whether they were taking Heinz's perspective or remaining objective.

Among participants who remained objective, reading that Heinz was wealthy caused people to judge stealing as less acceptable than those who read that Heinz was poor. Yet when participants engaged in perspective-taking, Heinz's wealth had no effect on judgments. This suggests that perspective-taking increases people's understanding of Heinz's situation, regardless of his financial circumstances.

Your results section should include a figure showing this interaction. A line or bar graph with wealth on the x-axis and separate lines for perspective-taking versus objective conditions helps readers visualize how these variables work together. Make sure to include error bars and clear labels.

After reporting the statistical results, explain what they mean for understanding moral judgment. The interaction suggests that people naturally consider someone's wealth when judging theft, but

perspective-taking might override this tendency by helping them understand the person's situation regardless of their economic circumstances. Consider how this connects to broader questions about when perspective-taking is most effective at changing moral judgments.

Remember that interactions can be complicated. Use clear language and concrete examples to help readers understand how your variables work together to influence moral judgments.

Consider what questions your results raise for future research: are there other factors that might moderate the effects of perspective-taking on moral judgment?

Your Turn: Exploring Moderators of Perspective-Taking

Now that you have worked through a factorial design, you can explore how other factors might influence when perspective-taking works best. The 2×2 design you just used can be adapted to examine many different moderators—variables that turn the effect of perspective taking on or off.

Consider which aspects of Heinz's situation might affect how people respond to perspective-taking. For instance, you might vary the relationship between Heinz and the person he's helping. Would perspective-taking have the same effect if Heinz was stealing medicine not for his wife (as in our example), but for a friend or a stranger? Do his actions perhaps become less understandable, and therefore less acceptable, as the distance between him and the person he's helping grows?

Using the factorial design, you could randomly assign participants to either take Heinz's perspective or remain objective, and to read about helping different targets. This design would allow you to examine whether perspective-taking is equally effective at increasing sympathy regardless of relationship or whether it works better for some relationships than others.

Alternatively, you might manipulate information about the company that designed the drug. Would perspective-taking increase people's acceptance of Heinz's behavior if the company was charging a reasonable markup for the drug versus price-gouging during a health crisis? Or what if people learned the company had spent hundreds of millions of dollars developing the drug versus acquiring the patent for the drug from another company? Does perspective taking work regardless of the reason why the drug costs so much?

Whatever you decide to manipulate, you can use the Qualtrics survey from our example as a template. All you need to do to conduct your own experiment is replace the wealth manipulation with a variable of your choice, while keeping the perspective-taking manipulation the same. Remember that factorial designs require larger samples, so you want to aim for about 50 participants per condition. In a 2×2 design that means gathering data from about 200 people.

Summary

Throughout this chapter, you have explored the powerful methodology of experimental research—the method that allows behavioral scientists to establish cause-and-effect relationships. Experiments can establish causality thanks to three essential elements: manipulation of an independent variable, random assignment of participants to conditions, and measurement of a dependent variable. You have learned that random assignment is the "magic" of experimentation. It ensures all potential third variables are equally distributed across conditions, allowing researchers to isolate the causal effect of their manipulation.

You have also explored several experimental approaches, beginning with a simple between-subjects design where different participants experience different conditions, such as in the God's views study. Then, you examined within-subjects designs, where the same participants experience all conditions in the study. These designs offer advantages like greater statistical power with fewer participants since each person serves as their own control, but they also present challenges like order effects that must be addressed through counterbalancing. The Heinz dilemma with different consequences (life-saving, pain relief, experimental drug) demonstrated this approach. Finally, you studied factorial designs that manipulate multiple variables simultaneously, allowing researchers to examine both main effects and interactions. Each of these designs helps behavioral scientists understand cause and effect relationships.

Through guided research projects, you gained hands-on experience implementing experimental manipulations like perspective-taking, randomly assigning participants to conditions using Qualtrics, analyzing experimental data with appropriate statistical tests (*t*-tests, chi-square, ANOVA), and interpreting and visualizing results from different experimental designs. These practical applications reinforced your understanding of experimental methodology.

As important as the hands-on activities are, don't forget the conceptual information you have learned in this chapter. In studying experiments, you have learned how researchers address the directionality and third-variable problems that limit correlational research. Experimental design can be remarkably simple yet powerful, as demonstrated by the God's views study that changed only the order of two questions. Given their simplicity, experiments are widely used beyond behavioral science in fields like medicine, agriculture, business, and policy, showing their versatility and value.

By mastering experimental techniques, you now have the tools to design studies that can move beyond describing relationships to understanding what causes what in human behavior—the ultimate goal of behavioral research. The ability to establish causality will serve as a foundation for your future work, whether you are conducting your own research or evaluating the claims made by others.

Preview

Chapter 8: AI Tools for Qualitative and Quantitative Research

What if you had a close friend or family member who fell deep into a conspiracy theory? Perhaps they believed the government was hiding evidence of aliens, that a secret cabal controlled world events, or that a popular health practice was harmful. Every time you shared contradicting evidence, you were met with resistance or dismissal.

How could you change this person's mind? Would it be best to ask questions and try to understand their perspective? Should you double down on facts? Should you share personal stories? What might persuade the person you care about to change their mind?

The challenge of persuading someone to change a deeply held but potentially harmful belief is a common experience. Whether it is a family member involved in a multilevel marketing scheme, a colleague who has embraced a fringe ideology, or a friend who rejects medical consensus, changing minds requires more than presenting facts. For that very reason, some researchers have recently wondered: could artificial intelligence offer a new approach to this problem?

In 2023, researchers from MIT and several other universities collaborated on a fascinating study. Their research examined whether personalized conversations with an AI interviewer could effectively reduce people's beliefs in conspiracy theories. The research team investigated this question through an online study involving over 2,100 participants who had strong beliefs in various conspiracy theories (Costello et al., 2025).

In the experiment, participants were asked to describe a conspiracy theory they found compelling and to explain their reasoning. Then, they engaged in a back-and-forth dialogue with an AI interviewer that was instructed to provide counterarguments and evidence challenging the person's beliefs. The results were impressive. The AI conversations reduced conspiracy beliefs by approximately 20% compared to a control group, and these effects persisted even when measured two months later. Even participants with deeply entrenched beliefs showed significant reductions in their conspiracy thinking after these personalized AI interactions.

Not long ago, this study would have been impossible. Having thousands of people engage in personalized conversations would have presented overwhelming logistical challenges. But today, using online participant recruitment platforms combined with AI-based survey tools, researchers can conduct personalized interviews on a massive scale.

In this chapter, we explore how AI is transforming survey research and facilitating qualitative and mixed methods research. In Module 8.1, we examine the fundamental changes AI is bringing to traditional survey methods. We discuss how AI allows dynamic, conversational approaches to survey design that can elicit deeper insights than static questions, and how these capabilities expand the toolkit available to behavioral scientists. Using specific examples, we show how AI-powered platforms can engage in flexible conversations and generate meaningful quantitative insights from open-ended data.

In Module 8.2, we explore a guided research project using the Heinz Dilemma to demonstrate how AI interviews can effectively combine qualitative and quantitative methods. We will walk through the practical aspects of setting up AI-interview, providing guidance to AI systems, analyzing the resulting data, and using AI tools for mixed methods analyses. This module will show how researchers can leverage AI to replicate the kind of in-depth qualitative research that traditionally required extensive human resources. Understanding these emerging technologies will likely be essential for any researcher looking to conduct innovative behavioral science in the coming years.

Chapter 8 Outline: AI-Survey Tools for Qualitative and Quantitative Research

Module 8.1: The AI Revolution in Survey Research: From Static Questions to Dynamic Conversations



Reading

Learn how artificial intelligence is transforming survey research by enabling dynamic, conversational data collection that goes beyond fixed-response questions.

- Discover how AI-survey tools allow for more interactive research studies
- Explore how researchers examine conversations for themes
- See how AI facilitates qualitative and mixed methods research

Module 8.2: AI and Mixed Methods Research: A Study of Moral Reasoning Using the Heinz Dilemma



Laboratory

Use the Heinz Dilemma to see how AI enables mixed-methods research at scale, combining the depth of qualitative insight with the efficiency of quantitative research.

- Walk through a guided project to see AI-survey tools in action
- Examine how AI helps with data analysis
- Learn about the role of the researcher when using AI tools

The AI Revolution in Survey Research: From Static Questions to Dynamic Conversations

Module 8.1

Learn how artificial intelligence is transforming survey research by enabling dynamic, conversational data collection that goes beyond fixed-response questions.

Artificial Intelligence (AI) is transforming many aspects of human life, from how people navigate traffic to how they discover music. In the social and behavioral sciences, AI is opening new possibilities, too. Back in Chapter 2, we described how ChatGPT's Data Analyst can conduct the statistical analyses covered in this book. In Chapter 4, we explored AI's ability to help create survey instruments. But among the areas where AI is having a particularly dramatic impact is in expanding the capabilities of survey platforms.

Traditional survey platforms are generally passive. They present participants with questions or statements and provide a fixed set of response choices. Yet with AI the research experience can be much more dynamic and interactive (e.g., Xiao et al., 2020; Chopra and Haaland, 2023; Wuttke et al., 2024).

At its core, AI refers to computer systems that perform tasks that typically require human intelligence. These systems can understand human language, engage in conversations, and ask follow-up questions, much like a human would. The ability to have flexible conversations with respondents opens many possibilities to expand research methodology. Let's examine a concrete example of how an AI-guided survey can enhance the depth of a survey.

Imagine we are conducting a study to understand why people drink alcohol. A typical quantitative study using the approaches you have learned about in previous chapters might ask respondents: "Why do you drink alcohol?" Participants would then be presented with a predetermined list of answer options (Figure 8.1) and asked to check all the reasons that apply from options like "to relax," "to socialize," and "to cope with stress."

How would the researcher come up with the response options in the first place? They could use previous research on the topic or they could generate their own reasons based on experience and logic. This type of research is often called **confirmatory** since it seeks to confirm and to quantify an already existing set of hypothesized reasons for the behavior in question. By design, a confirmatory approach is constrained to categories defined by the researcher.

Traditional Survey Approach

In traditional survey research, questions about alcohol consumption typically use closed-ended formats with predetermined response options.

Q: Why do you drink alcohol? (Select all that apply)

- To relax
- To socialize with friends
- To cope with stress
- I enjoy the taste
- To celebrate special occasions
- It helps me sleep
- Other (please specify): _____

Figure 8.1 A traditional fixed-response question that aims to understand the reasons people drink alcohol.

However, what if the goal was to identify novel reasons for drinking alcohol that previous studies may have missed? Or, what if the research on a topic is sparse and there is not a large literature to lean on? In such cases, researchers can adopt an **exploratory** approach. Instead of presenting predetermined answer options, a researcher can conduct an exploratory study in which they engage respondents in a natural conversation. Respondents can be asked to describe their reasons for drinking in an open-ended format, where they are free to say whatever they want.

As illustrated in Figure 8.2, the conversation might begin with a simple question about whether the person drinks alcohol. Then, it can explore their reasons for drinking or abstaining. When a respondent mentions drinking "when out with friends," the AI can inquire further about how alcohol functions in social situations. This conversation continues, with the AI asking follow-up questions based on the specific content of each response, allowing respondents to express themselves in their own words.

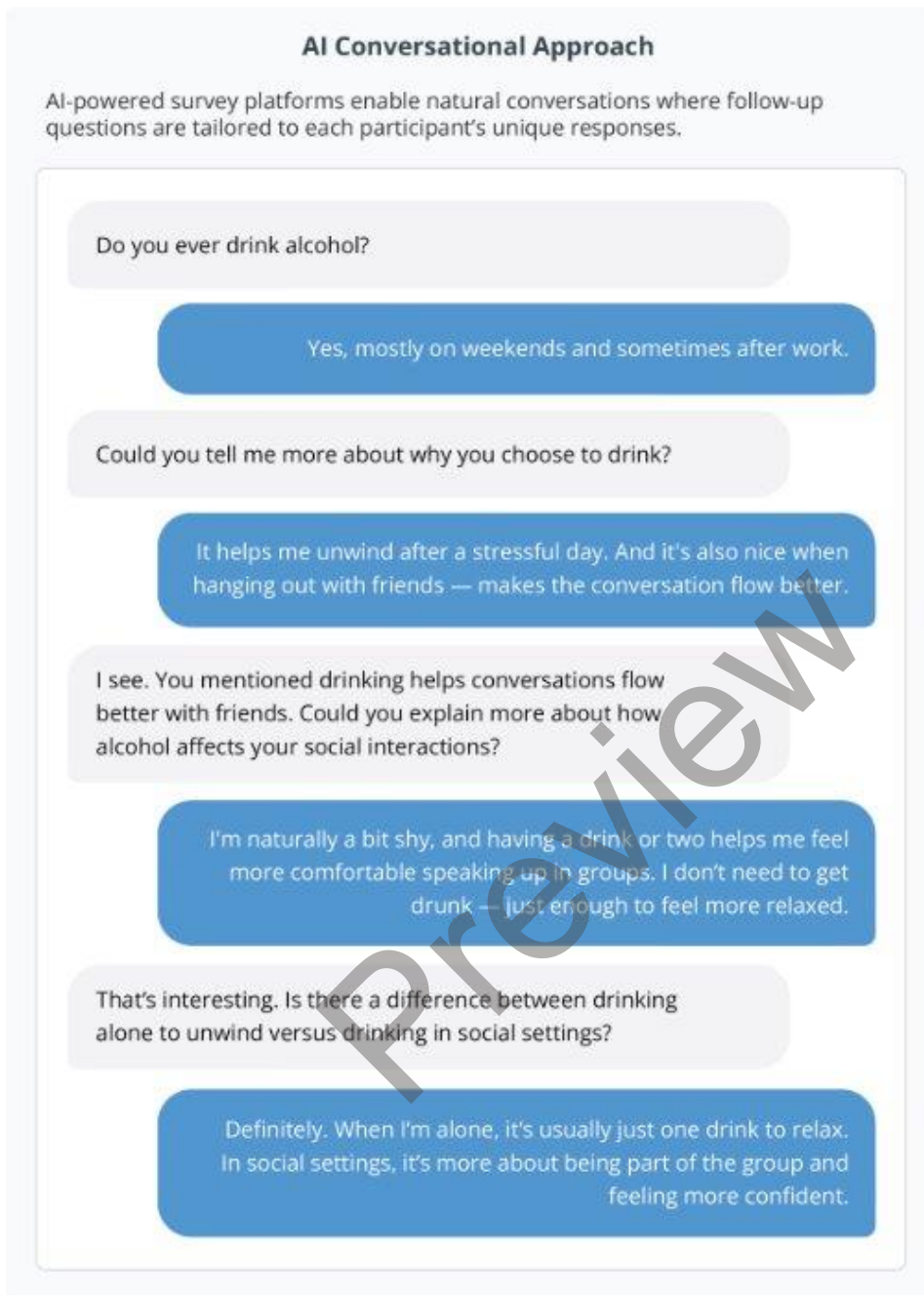


Figure 8.2. A hypothetical conversation with a respondent that aims to uncover the WHY behind drinking behavior in open-ended conversation.

The example above is a simplified scenario. To illustrate how an actual AI survey can conduct a project of this nature, we gathered responses from 100 people on Connect and asked them about their alcohol consumption. The AI-powered platform we used is called Engage. Engage

generally functions like Qualtrics and can even be used to upload the Qualtrics .qsf files we have used throughout this book. The key difference, however, is that Enage has additional AI interviewing capabilities.

The Enage interviewer had a conversation with each respondent about their reasons for drinking alcohol and encouraged elaboration without leading respondents toward any particular answer. You can explore the study, all one hundred conversations, and the results, by following this link to the Enage platform: <https://bit.ly/43dyOVt>. You can login with the CloudResearch credentials you have used previously or create a new account.

After completing the interviews, the AI analyzed the text data and listed the reasons participants provided for drinking. Figure 8.3 shows the eleven reasons Enage identified from the transcripts.



Figure 8.3. Motivational themes for drinking alcohol extracted by AI from AI-conducted conversational interviews.

The analysis revealed a variety of motivations for drinking including social facilitation, where people described alcohol as a "social lubricant" that helps to interact with others (51%), drinking

simply because people enjoyed the experience (47%), coping with stress (43%), and drinking for relaxation (40%).

Some, perhaps, less obvious motivations emerged as well. Thirteen percent of people mentioned drinking to enhance experiences like food and music, while 11% noted drinking to enhance the flavor of meals. The least common motivation, mentioned by 4% of people, was feeling obligated to drink despite having no desire to do so.

Of course, the reasons for drinking alcohol in this sample are not comprehensive because the sample was relatively small. However, when conducted with thousands of participants and across numerous demographic, cultural, and globally diverse groups, this kind of research can paint a rich tapestry of the reasons people have for drinking alcohol, and it can do so faster and with less effort than perhaps any other method. Such research can have numerous practical applications. With enough respondents in the sample, it becomes possible to examine how motivations for drinking differ among groups, such as men and women, teenagers and adults, and people in different socio-economic groups. A detailed analysis of drinking motivations can inform public health campaigns, which could be tailored to different people depending on their motivations. And market research campaigns could use such insights to develop and market products that better align with consumers' motivations.

What this example demonstrates is that AI-powered conversational research can enhance the study of human behavior and motivation. It differs from the quantitative approach in several ways (Figure 8.4).

Traditional survey methods typically predetermine the possible reasons for drinking alcohol and participants must select from these fixed options. In contrast, AI-powered conversational surveys adopt a more flexible and exploratory approach where participants express their motivations in their own words through natural dialogue. The AI then identifies patterns and emergent categories from the conversations. By enabling natural conversations with hundreds or even thousands of participants to be conducted and analyzed in a matter of minutes researchers can uncover the “why” behind people’s behavior and motivations in novel ways that were harder to implement with more traditional methods.

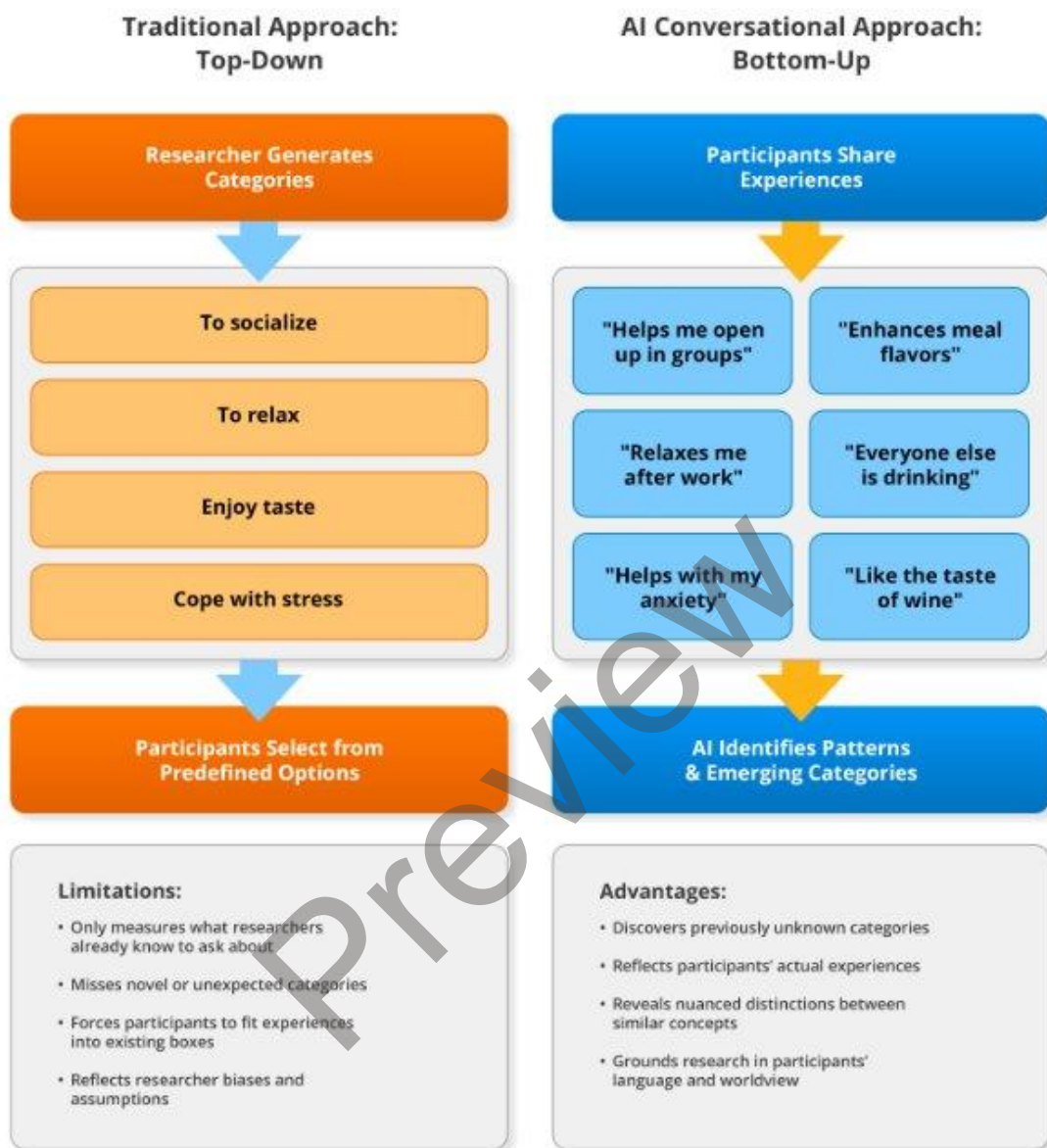


Figure 8.4. A diagram of quantitative and qualitative approaches, with a focus on the advantages of qualitative methods.

Qualitative, Mixed Methods, and Quantitative Research

The study discussed above is an example of **qualitative research**. While it used AI to conduct the interviews and analyze the data, traditional qualitative research is done by people and is a common methodology across many domains of behavioral science. In fields such as anthropology, sociology, and social psychology researchers use qualitative methods to answer a wide range of research questions.

In sociology, for instance, entire books have been written about wide-ranging questions in human life based on qualitative interviews. For example, in a classic best-selling book, *The Seasons of a Man's Life* (1978), Daniel Levinson conducted in-depth qualitative interviews with people across a variety of ages. From these interviews, he discovered patterns in how adults develop and change throughout their lives. By analyzing people's stories, Levinson identified and described distinct periods in adult development, each with its own specific challenges.

Qualitative research has also produced several other fascinating books. In *Hooking Up: Sex, Dating, and Relationships on Campus*, Kathleen Bogle (2008) described the hookup and dating culture on contemporary college campuses. Elliot Liebow (1993) described the daily struggles of women who lived on the streets in *Tell Them Who I Am: The Lives of Homeless Women*. And, in *The Tenants of East Harlem* (2006), Russell Leigh Sharman describes the realities of changing inner city life in a New York City neighborhood.

Beyond interviewing people, qualitative research also has direct applications to quantitative research. For example, using a method called **cognitive interviewing**, qualitative interviews are used for questionnaire development. In a cognitive interview, a researcher asks people to "think aloud" as they answer specific questions within a questionnaire, sharing everything that goes through their minds. Such question development techniques are invaluable for validating survey questions to make sure people understand the questions as they were meant to be understood. Without such validation, questions can be misinterpreted by respondents. For example, in one study, people were asked, "How many times have you visited a doctor in the past year?" and many were unsure whether to count phone consultations or visits to nurse practitioners (Willis, 2004). Without the opportunity to clarify, the researchers might never have known their questions were being interpreted inconsistently.

The Role of AI in Qualitative and Mixed Methods Research

Despite the usefulness of qualitative research, its implementation has traditionally faced major practical challenges. The first challenge involves the time and resources required to conduct open-ended interviews. A skilled researcher might complete three or four interviews in a day. This means a study with 30 participants, a typical sample size for qualitative research, would take two or three weeks just to collect the data.

The second challenge comes after the interviews are complete: making sense of the data. Imagine sitting down with 50 interviews, each filled with rich descriptions of people's experiences and thoughts. Researchers need to read through each interview multiple times: first to familiarize themselves with the content, then to identify patterns, and finally to carefully code each response. These analyses can take months.

The time required to conduct qualitative research explains why these studies typically have small sample sizes. Whereas a quantitative study might include hundreds or even thousands of participants, qualitative studies usually work with 20-30 participants (see Yin, 2016). And, even this small number generates an enormous amount of information to analyze.

This is where AI is changing qualitative research in two important ways. First, AI can conduct in-depth interviews with many participants simultaneously (see Austin et al, 2025; Costello et al., 2025). Like a skilled human interviewer, AI can ask follow-up questions when responses are brief or unclear, probe deeper when participants mention something interesting, and maintain a natural flow of conversation. This means researchers can conduct qualitative interviews with hundreds of participants in the time it would traditionally take to interview just one person (see Costello et al, 2025).

Second, AI can analyze large amounts of qualitative data quickly. Researchers can use AI to identify patterns across hundreds of interviews, either by applying existing theories (like coding responses into pre-defined categories) or by discovering new patterns they had not anticipated. While human researchers still need to guide this process and interpret the findings, AI can dramatically speed up the analysis. The few studies that have examined the reliability of AI for qualitative data analysis have generally reported promising results (e.g. Austin et al., 2024)

In the next section, we will conduct a guided research project with the Heinz Dilemma to help illustrate how AI interviews can combine qualitative and quantitative methods.

Preview

AI and Mixed Methods Research: A Study of Moral Reasoning Using the Heinz Dilemma

Module 8.2

Use the Heinz Dilemma to see how AI enables mixed-methods research at scale, combining the depth of qualitative insight with the efficiency of quantitative research.

To understand AI interviews, let's return to the Heinz dilemma used in earlier chapters. While we have used quantitative methods to examine the Heinz dilemma, Kohlberg developed his theory of moral development through intense qualitative research. He spent years conducting in-depth interviews, following adolescents as they grew up and carefully analyzing how their moral reasoning changed over time (Kohlberg, 1984).

In his research, Kohlberg presented moral dilemmas like the story of Heinz. Then, he interviewed participants about what Heinz should do and why. Through careful analysis of these conversations, he noticed patterns in how people justified their moral decisions. Some people focused on avoiding punishment, others on following rules, and still others on abstract principles of justice. These observations led him to identify six distinct stages of moral development.

In previous chapters, we conducted descriptive research to understand how many people think it is acceptable for Heinz to steal the drug, correlational studies to examine which variables might predict this decision, and experiments to test whether taking Heinz's perspective causes people to be more accepting of his decision to steal. Now, we will use AI to mimic Kohlberg's qualitative approach.

We will attempt to replicate the results of Kohlberg's interviews by exploring how people think through this moral decision in their own words. What factors do people consider? How do they weigh different principles against each other? How do they justify their conclusions, and what do their reasons reveal about moral judgment?

As with other guided research projects in this book, we have conducted the study, analyzed the data, and made the results available to you. You can find them at this link:

<https://bit.ly/43dyOVt>. After reviewing the study, we encourage you to replicate the project with friends or family who are willing to take ten minutes to complete the study, student participants on SONA, or with participants from Connect. You only need about 20 respondents to start

exploring AI surveys. For a more comprehensive analysis, we recommend 100 to 200 respondents

For our qualitative study of Kohlberg's moral stages, we used Engage, which was specifically designed for qualitative and mixed-methods research. Like traditional platforms, Engage allows researchers to create a survey link to share via email or participant recruitment platforms like Connect. Both the Engage and Connect are made by CloudResearch so they can be integrated seamlessly. The study has already been created on Engage, where you can work with it similarly to how you have worked with Qualtrics on previous assignments.

AI-Conducted Interviews: A Conversation About the Heinz Dilemma

To examine the stages of moral reasoning in the Heinz Dilemma, we conducted 100 interviews with participants who were recruited from Connect. In our study, we began with a simple yes/no question: "Should Heinz have broken into the pharmacy to steal the drug for his wife?" But the real insights come from understanding why people answer the way they do—exactly what Kohlberg was after when developing his theory of moral development.

To receive those answers, we first needed to tell the AI how to have the conversation. In Engage, this takes the form of an interview guide—similar to what a human interviewer would use. You can see the guide and how it follows a quantitative question in Figure 8.5. We instructed the AI to probe deeper with follow-up questions that would encourage participants to share their reasoning process. The interview guide told the AI to ask open-ended questions about specific factors influencing the participant's decision, to explore emotional aspects of their reasoning, and to investigate how they viewed the relationship between the law and morality.

2. Multi-Choice Select One *i*

Engage Off *i*

Should Heinz have broken into the pharmacy to steal the drug for his wife?

Yes

No

What factors influence your decision-making in this scenario?

Anchor question to start the open-ended interview

Follow-up Instructions* Engage *i*

Tell Engage what your objectives are and how to follow up on this question

Continue asking open-ended questions, to understand the specific factors that influence the participant's decision-making in the context of the Heinz Dilemma.

Ask open-ended follow-up questions that encourage participants to share personal insights, emotions, and detailed explanations. Probe for specific influences on their stance regarding whether Heinz should steal the drug to save his wife: What specific aspects of the situation most strongly shaped your view on that? If you mentioned emotional factors, can you share any personal experiences or beliefs that connect to how you feel about Heinz's situation? If you bring up legal or societal norms, how do you see the relationship between legality and morality in this case?

Discussion guide given to the AI-interviewer to guide the open-ended conversation

2036 / 5000

Figure 8.5. Engage allows researchers to follow close-ended questions with an open-ended item that asks participants to explain why they chose a particular answer.

A sample conversation is shown in Figure 8.6. Notice how the AI acknowledges the participant's key points and asks for elaboration on specific aspects of the conversation. Having set up this study on Engage and then having launched it on Connect, we were able to collect over 100 in-depth qualitative interviews in about 1 hour. You can explore all the conversations by following the survey link provided above.

Should Heinz have broken into the pharmacy to steal the drug for his wife?

Yes
 No

Thank you for your response. Can you explain why you believe Heinz should have stolen the drug? What influenced your decision?

This is a life or death situation. Heinz did everything that was necessary the legal way to get the funds to buy this drug for his wife, and he still came short. If he went straight to stealing without exhausting all of his other options, then I would not have selected yes, but his wife is about to die. Also, the only person really facing any loss in this situation is the company, and they honestly kind of deserve it after selling a drug that could save lives for \$20k when it only cost \$2k to produce. If I was Heinz and the person that I love dearly was about to die, I would also consider risking going to jail to save them.

I appreciate your thoughtful response. You've emphasized factors such as the desperate situation Heinz was in and the perceived unfairness in the pricing of the medication.

Looking into the moral aspects, do you think this situation justifies the act of stealing, generally considered as wrong? If yes, under what circumstances stealing can be considered as a justifiable act according to you?

It is somewhat justifiable. The only times I think stealing could even possibly be considered justifiable is when the stealing is done against big, selfish corporations. If the stealing was done from individual people or small businesses, then I think it would be wrong. Also, in this specific situation, it was life or death. I think stealing would be better than letting his wife die.

Thank you for elaborating on your perspective. Your response suggests that you view the morality of stealing within a broader social and economic context, specifically whether the entity being stolen from is perceived as exploiting others.

To dive a bit deeper into this: Do you hold the pharmaceutical company responsible for this unfortunate situation? To what extent do you think they played a part in it? Do you think they could have done something differently to prevent it?

Yes. Absolutely. They are absolutely responsible. Pharmaceutical companies should not be allowed to mark up drugs that could potentially save lives to that extent. Why should healthcare only be accessible to those who can afford it? I feel like it is a basic human right to be able to have access to these resources. The company acted selfishly and wanted to exploit sick people to make a profit. If they wanted to mark up the prices to make some money, I think there should be regulation specifically for pharmaceutical companies on how much they could mark up their products.

Figure 8.6. A sample interview between an AI interviewer and a human participant.

AI-Based Approaches to Developing a Coding System

After conducting interviews about the Heinz dilemma, the next challenge was to analyze how participants reasoned about their decision and to group the responses into categories. In our first analysis, we aimed to categorize responses according to Kohlberg's six stages of moral development.

The first step in this analysis was to provide the AI with a clear definition of each category. For our study, we gave detailed descriptions of all six of Kohlberg's stages, including examples of how people at each stage might reason about the Heinz dilemma. For instance, we described Stage 1 responses as focused on avoiding punishment ("Heinz shouldn't steal because he could go to jail"), while Stage 6 responses appealed to fundamental moral principles ("Saving a life is a higher moral obligation than respecting property rights"). These descriptions can be seen in Figure 8.7.

After providing the AI with these detailed descriptions, it analyzed each participant's responses and tagged them according to Kohlberg's stages of moral development. Let's look at some examples to see how the AI identified and categorized different stages of moral reasoning.

Figure 8.8 shows how the AI tagged a participant's response as Stage 2 (Self-Interest Orientation). When the AI analyzes responses, it provides its reasoning for each decision. This allows researchers to verify whether they agree with the AI's categorization and understand exactly why the AI made each decision. For this respondent, the AI-provided explanation was: "This quote highlights drastic measures taken in a situation of survival, which aligns with self-interest as it focuses on doing what one can to survive. The reasoning is about individual need and survival rather than any social or ethical obligation." This matches our definition of Stage 2 reasoning, which emphasizes practical consequences and reciprocal thinking.

Tag ▼ ⓘ Off Mutually Exclusive ⓘ 🗑️

Tag participants as 'Stage 1: Obedience and Punishment Orientation' if their moral reasoning is based on avoiding punishment. For example, their reasons might be "Heinz shouldn't steal the drug because he could get caught and go to jail."

Tag participants as 'Stage 2: Self-Interest Orientation (Instrumental Relativist)' if their Morality is defined by what is in one's best interest or by a reciprocal exchange. For example, their reasons might be "Heinz should steal the drug because he needs it to save his wife, and later he can pay the druggist back."

Tag participants as 'Stage 3: Interpersonal Accord and Conformity (Good Boy/Good Girl Orientation)' if their Moral reasoning is influenced by social roles and expectations, with an emphasis on relationships and approval. For example, their reasons might be "Heinz should steal the drug because he loves his wife, and a good."

Stage 1: Obedience and Punishment Orientation ×

Stage 2: Self-Interest Orientation ×

Stage 3: Interpersonal Accord and Conformity ×

Stage 1: Obedience and Punishment Orientation
Stage 2: Self-Interest Orientation
Stage 3: Interpersonal Accord and Conformity

Description of 'Stage 1' of moral reasoning given to AI for categorizing qualitative responses

Figure 8.7. Instructions provided to the AI survey system for analyzing qualitative data.

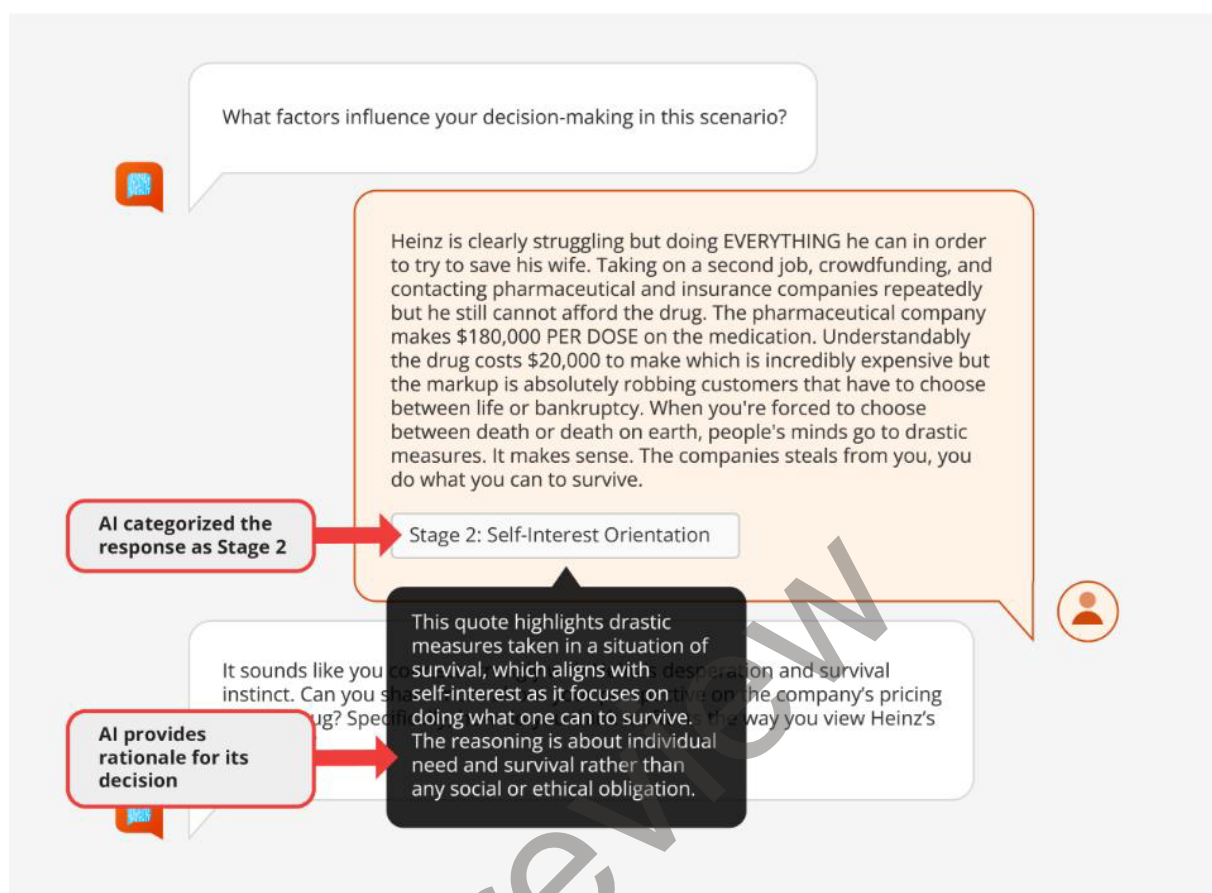


Figure 8.8. A participant's response was categorized as Stage 2 of moral reasoning. The AI places the response into Stage 2 and provides a rationale for its decision.

Figure 8.9 shows how another respondents' conversation was categorized as indicating Stage 4 reasoning. The AI's explanation was: "This quote aligns with Stage 4: Authority and Social Order Maintaining Orientation because it emphasizes the importance of following the law and finding legal means to solve problems. The speaker suggests that stealing could set a bad precedent, which underscores a concern for societal rules and order."

This categorization reflects how the participant focuses on maintaining social order ("slippery slope") and following legal procedures rather than considering either individual needs (Stage 2) or universal principles (Stage 6).

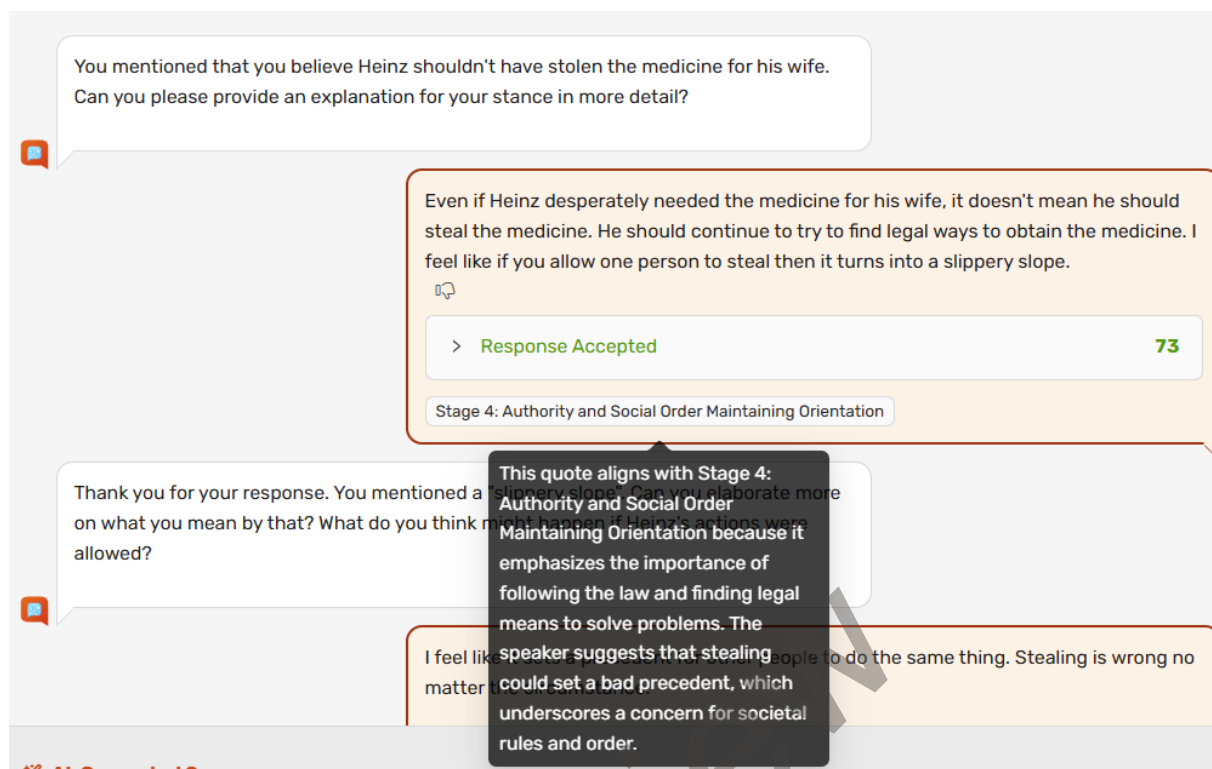


Figure 8.9. A participant's response was categorized as Stage 4 of moral reasoning. The AI places the response into Stage 4 and provides a rationale for its decision.

AI Tools for Mixed Methods Analyses

While the AI's ability to categorize individual responses uncovers how different people reason about moral decisions, it can also be used to identify broader patterns across responses. This is where AI enables a true **mixed-methods** approach. Broadly defined, mixed-methods is any technique that combines qualitative research with quantitative analysis (see Curry and Nunez-Smith, 2015). One way to bring a mixed methods approach to the current study is to simply count the number of times each moral reasoning stage was detected.

After counting how often each stage is observed, we can see the distribution of different types of moral reasoning in the sample, as shown in Figure 8.10. Stage 5 (Social Contract Orientation) appeared most frequently, showing up in 62% of responses. This was followed by Stage 2 (Self-Interest Orientation) at 45% and Stage 4 (Authority and Social Order) at 38%. Because

participants often used multiple types of moral reasoning in their responses, the percentages sum to more than 100%.

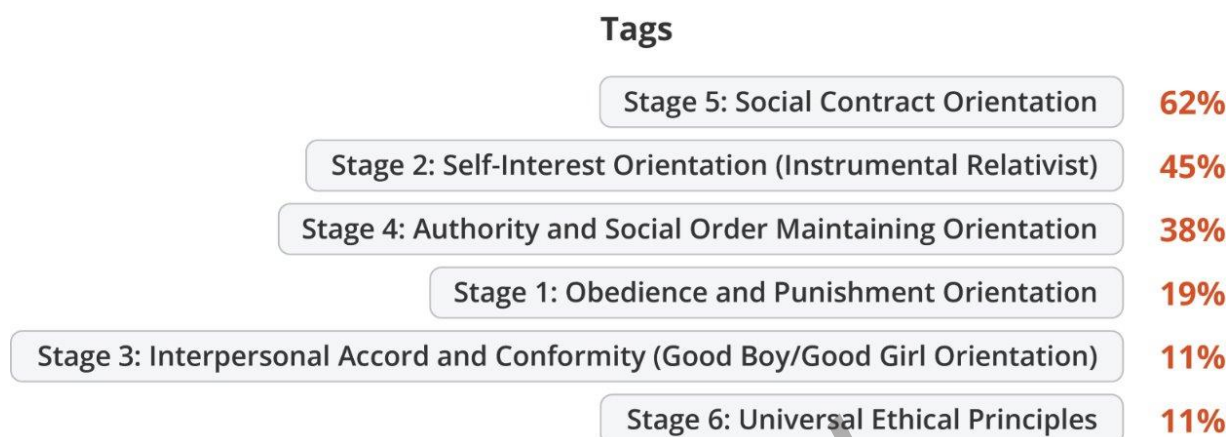


Figure 8.10. The AI can tag how often different themes are mentioned by participants.

This kind of quantitative summary can reveal patterns that might not appear when looking at individual responses. For instance, we can see that while participants commonly appealed to social contracts (Stage 5) and practical consequences (Stage 2), relatively few people (11%) reasoned at the highest level of moral reasoning, universal ethical principles (Stage 6).

Using AI Tools for Thematic Analyses and Discovering Unexpected Patterns

Using predefined categories like Kohlberg's stages or moral development to identify specific patterns in data is an example of top-down data analysis. But some of the most valuable insights in qualitative research come from discovering unexpected patterns in the data—what is often called a bottom-up approach. After using AI to categorize responses into Kohlberg's stages, we asked it to analyze the same conversations with fresh eyes, looking for themes and patterns we had not anticipated. This process is part of **exploratory analysis**.

While the goal of an exploratory analysis was not to critique Kohlberg's theory—there is much more depth to his system than we can cover in this chapter—this example shows how AI can be used to explore qualitative data in new ways. Similar analyses could be done to discover unexpected patterns in any research area. For example, in interviews about career choices, we might find novel decision-making factors. In studies of consumer behavior, we might discover

new motivations for people's purchasing decisions. And in research on student learning, we might explore the strategies people use to study. In each of these areas, an exploratory analysis might reveal things that other methods would not.

The Researcher's Role in AI-Driven Research

As transformative as AI is, it is important to emphasize the complementary relationship between human and AI when using these tools. While AI can interview people and process thousands of responses in minutes, the researcher's expertise remains central. Researchers must determine which patterns are meaningful, craft queries to explore these patterns, interpret the significance of what was found, and connect the findings to broader theoretical frameworks. AI accelerates the analysis process, but it does not replace the need for human judgment or expertise.

The key advantage of using AI is that exploratory analyses can be done almost instantly. What once might have taken months can now be accomplished with a few clicks, allowing researchers to conduct qualitative analyses at the speed of quantitative research. Indeed, by enabling natural conversations to be conducted at scale, AI-based methods help bridge the gap between quantitative and qualitative approaches. As AI systems become more sophisticated in their conversational abilities, researchers may find it increasingly useful to conduct studies that combine the best aspects of structured surveys, in-depth interviews, and large-sample statistical analysis to expand our understanding of human behavior across countless domains (see Austin et al., 2025).

Summary

In this chapter, we explored how artificial intelligence is transforming research in the behavioral sciences. From enhancing traditional surveys with dynamic, conversational approaches to enabling qualitative research at unprecedented scales, AI is expanding the toolkit available to researchers in ways that were unimaginable just a few years ago. The MIT study on changing conspiracy beliefs that opened the chapter highlighted how AI can personalize interactions with thousands of participants simultaneously, while our exploration of the Heinz dilemma demonstrated how AI can both conduct in-depth interviews and analyze the volume of data these conversations produce.

The capabilities of AI-driven survey platforms bridge the traditional divide between qualitative and quantitative research. Using platforms like Engage, researchers can now conduct hundreds of in-depth interviews in the time it would traditionally take to complete just one. Moreover, the ability to analyze these conversations both through existing theoretical frameworks (like Kohlberg's stages) and through exploratory analyses offers a powerful combination of confirmatory and discovery-oriented research. The result is a more nuanced and comprehensive understanding of human behavior, motivation, and decision-making. And these tools are only going to improve over time, which brings us to a good place to end Part I of this book: at the intersection of traditional research methods and the newest tools for behavioral research.

In Part I of this book, you have built a strong foundation of knowledge about behavioral research. You learned about descriptive, correlational, experimental, and mixed methods designs; you learned about the tools researchers use to facilitate modern research; you learned how to find, develop, and validate measurement instruments; you explored several techniques of statistical analysis; and you saw how AI can enhance traditional approaches to research. This knowledge forms the bedrock of behavioral science.

In addition to gathering knowledge, you had several opportunities to conduct your own research. By walking through guided projects before testing your own ideas, you have developed one of the most important qualities in learning to conduct behavioral research: experience. This experience should help you tackle future projects with confidence.

In Part II, you will build upon the fundamentals you have acquired and learn to apply them in the context of online research. In particular, you will discover how to design studies that take advantage of the unique capabilities of digital platforms while navigating their potential challenges. Whether you are conducting a simple survey, a complex experiment, or a mixed-methods study combining qualitative and quantitative approaches, the principles and practices covered in Part II will equip you to implement your research effectively in the online environment.

In many ways, the evolution of research methodologies parallels broader technological transformations. Just as AI has moved from science fiction to everyday reality, research methods have progressed from paper-and-pencil surveys to sophisticated digital interactions. The future of

behavioral science lies at the intersection of classic methodological rigor and these technological innovations. By mastering both the fundamentals covered in Part I and the strategies for implementation covered in Part II, you will be equipped to conduct research that is not only scientifically sound but also leverages the full potential of contemporary tools. This combination will allow you to address nearly any question that sparks your curiosity.

Preview

Appendix: Instructions for Writing Up the Results of a Study

After completing a research project, it's time to communicate what you found. Use the report guidelines below to write up the results of each independent project you conduct in your course. Each report should be approximately 2-3 pages long and include the following sections.

Abstract - Do this last!

Begin your report with a brief abstract of 150-200 words that provides a complete overview of the study. The abstract should include the purpose of your research, how you conducted it, your key findings, and the main conclusion. Think of this as a condensed version of your entire paper that someone could quickly read to understand what you did and what you found.

Your abstract should touch on your research question, the variables you examined, your sample, the method of data collection, the statistical approach, key results including correlation coefficients, and the significance of your findings. Write this section last, after you've completed the rest of your report.

Here is an example abstract: "This study examined the relationship between mindfulness practice and perceived stress levels among college students. Participants (N = 124) completed online measures assessing their frequency of mindfulness practice and perceived stress. Results revealed a significant negative correlation between mindfulness practice and stress ($r = -.41$, $p < .001$). These findings suggest that regular mindfulness practice may be associated with lower stress levels in college students, though the cross-sectional nature of this study prevents causal conclusions. Future research should examine these variables longitudinally to better understand how mindfulness practice affects stress over time."

Introduction

Begin with a clear statement of your research question and hypothesis. What relationship were you investigating? What did you predict you would find, and why? This section should be brief but specific, focusing on the variables you measured and the relationship you expected to find between them.

Literature review

Indicate whether your study is:

- A replication of a previously established effect
- A modification of a known relationship
- An investigation of a novel relationship

To determine this, conduct a literature search using Google Scholar, as you learned in Chapter 2. This will help you contextualize your study within existing research. Make sure to reference any key studies that inform your hypothesis or relate closely to your research question.

Method

Describe how you conducted your study in straightforward terms. Include information about the participants: How many people participated in your study? What were their key demographics (average age range, gender distribution)? Then, describe the measures you used. Which variables did you measure and how? Describe each scale or measurement tool, including well-formatted tables that list the items in each measurement scale you used. Finally, describe the procedure. How did you collect your data? Was it through Connect, student participants, or another source? What software was used for stimulus presentation (Qualtrics)?

This section should be factual and concise. Provide enough information that someone could understand what you did without overwhelming them with details.

Results

This is where you present what you found. Focus on the key statistical findings that address your research question. If your study is descriptive, present statistics that describe people's responses. If the study is correlational, report the correlation coefficient(s) between your main variables, noting both the strength and direction of the relationship. And, if the study is experimental, report the tests that allow you to determine if there was an effect of your independent variable(s).

Mention whether your results were statistically significant. Describe any interesting patterns in your data, such as unexpected relationships between variables. Include a figure that visually represents your main finding (a scatter plot or bar graph), and then include all relevant SPSS syntax and output in an **Appendix**. This is like showing proof of your work.

Remember to present your results in plain language alongside the statistical information. For example: “A Pearson correlation analysis was conducted to examine the relationship between anxiety and depression. There was a significant positive correlation between anxiety (as measured by the GAD-7) and depression (as measured by the PHQ-9), $r(524) = .82, p < .001$. These results suggest that anxiety and depression are strongly related. Specifically, the results suggest that people who experience higher levels of anxiety also tend to experience higher levels of depression.”

Discussion

Interpret what your findings mean in about a paragraph or two. Did your results support your hypothesis? Why or why not? How do your findings connect to previous research or theories? What might explain the patterns you observed?

Also, briefly mention at least one limitation of your study and one idea for future research that could build on your findings.

Conclusion

End with a brief statement (2-3 sentences) summarizing the main takeaway from your study and why it matters.

Formatting Guidelines

- Write in clear, non-technical language that anyone with basic knowledge of behavioral science could understand.
- Use proper APA-style formatting for reporting statistics.
- Include one figure that visually represents your main finding.
- Double-space your document with 1-inch margins.
- Include a title that captures the essence of your study.

Reference section

Include a reference section that lists all the sources you cited in your report. Format your references according to APA style guidelines. At minimum, you should include references for:

1. Any previous research you mentioned when contextualizing your study

2. The sources of any established measures or scales you used
3. Any theories or frameworks you referenced in your discussion

For example:

Burke, M., & Kraut, R. E. (2016). The relationship between Facebook use and well-being depends on communication type and tie strength. *Journal of Computer-Mediated Communication*, 21(4), 265-281.

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior*, 24(4), 385-396.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, 37(6), 504-528.

Preview

Part II: Conducting Online Research

Preview

Chapter 9: Finding Participants

In the summer of 2024, OpenAI—the company that created ChatGPT—collaborated with the MIT Media Lab to conduct a study that would have been nearly impossible just a few years before. The study examined how frequent interactions with AI chatbots affect human psychology.

In the study, researchers followed nearly 1,000 people for a month. Each day, participants completed various tasks and interacted with ChatGPT. Then, at the end of the study, the researchers measured how these interactions affected people's feelings of loneliness and their emotional dependence on AI. One of the study's main findings was that interacting with ChatGPT reduced people's feelings of loneliness, but at the cost of becoming more dependent on AI (Phang et al., 2025). What makes this study interesting, however, isn't only its findings. It is the fact it happened at all.

Just a few decades ago, following 1,000 people every day for a month would have presented overwhelming logistical challenges. Researchers would have struggled to track the behaviors of so many people and prevent most of them from dropping out of the study. The project would have required enormous amounts of time and money, and even then, it would have faced long odds of success. Times have changed.

Today, online tools make it much easier to find research participants and to conduct a wide range of behavioral studies. In fact, this study was conducted on CloudResearch Connect—the same platform that was used to collect data for all the projects in Part I of this book. Using online platforms like Connect, researchers can find participants, communicate with them throughout the study, offer payments and bonuses that incentivize participation, and monitor each person's progress over the course of the study. All these activities can be accomplished with just a fraction of the time and resources that offline studies require, which is why online research has transformed behavioral science over the last twenty years.

In this chapter, we describe the options for online participant recruitment in the behavioral sciences. In Module 9.1, you will learn about the history of participant recruitment in the

behavioral sciences and how online tools have transformed research in the last two decades. We will describe the different sources researchers use to find online participants beginning with the largest part of the online ecosystem: market research panels. Each year, these panels facilitate approximately 5 billion surveys. We will describe how they operate, which industries they power, and the tradeoffs they present to researchers. Then, we will take a closer look at researcher-centric online platforms like Connect and Mechanical Turk and explain why these platforms host most academic studies.

After exploring where researchers find participants, Module 9.2 discusses sampling and issues of representativeness in online samples. We will describe probability and non-probability sampling and the fit-for-purpose framework, which is useful for making decisions about which sources of participants are the right fit for specific projects. By the end of the chapter, you will understand the different ways to recruit participants online, the advantages of each approach, and how to choose the right participant source for your research question. Understanding these issues is an important part of successfully conducting online research.

Chapter 9 Outline: Finding Participants Online

Module 9.1: Options for Online Participant Recruitment



Reading

Learn about the options for recruiting participants online

- Read about the history of finding participants in behavioral research
- Learn about market research panels, the biggest part of the online ecosystem
- Explore the strengths and uses of crowdsourcing platforms

Module 9.2: Representativeness in Online Sampling



Reading

Explore issues of sampling in online research

- Learn about different types of sampling in behavioral research
- Understand why behavioral researchers often use non-probability samples and how they make those samples more representative
- Explore the fit for purpose framework to sampling

Options for Online Participant Recruitment

Module 9.1

Learn about the options for recruiting participants online

A Brief History of Finding Participants

In the summer of 1961, Stanley Milgram began work on what would become one of the most famous studies in the history of psychology. Driven by a pressing question—why do people obey unethical orders?—he designed a study in which ordinary people were asked to deliver increasingly powerful electric shocks to a stranger whenever that person gave incorrect answers on a memory test (the stranger was actually an actor).

In a surprise to many, eighty percent of participants delivered shocks up to 150-volts—a point at which the person receiving the shocks screamed, complained of a new condition, and began asking to be released from the study. More than 60% of participants continued all the way to the maximum shock of 450-volts, an area labeled "DON'T PUSH THIS LEVER TO AVOID SEVERE SHOCK: XXX" on the control panel and long past when the stranger in the other room had stopped responding (Milgram, 1963).

While the results of Milgram's study have been discussed and debated for decades (e.g., Benjamin & Simpson, 2009; Gilbert, 1981; Griggs, 2017), less attention has been paid to how he found participants. To recruit people, Milgram placed an ad in the local newspaper (Figure 9.1). The ad sought men between the ages of 20 and 50 who were employed in various occupations (later studies included women). Anyone interested in participating was asked to mail a slip of paper with their contact information, demographic data, and an indication of when they could participate. After receiving the replies, Milgram's team had to call each prospective participant and schedule a time for their participation. They also paid people for their time. Milgram offered each participant \$4 plus carfare, about \$50 in today's dollars.

This method of community-based recruitment was not unique to Milgram. In fact, it was a standard technique for finding people to participate in research at the time. In addition to running ads in the local newspaper, researchers would often post fliers on bulletin boards, recruit by word of mouth, or send direct mail to potential participants. All these recruitment options were slow

and limiting. Recruiting enough people for one study could take several months and researchers were restricted in the diversity of people they could reach.

**WE WILL PAY YOU \$4.00 FOR
ONE HOUR OF YOUR TIME**

Persons Needed for a Study of Memory

*We will pay five hundred New Haven men to help us complete a scientific study of memory and learning. The study is being done at Yale University.

*Each person who participates will be paid \$4.00 (plus 50c carfare) for approximately 1 hour's time. We need you for only one hour; there are no further obligations. You may choose the time you would like to come (evenings, weekdays, or weekends).

***No special training, education, or experience is needed. We want:**

Factory workers	Businessmen	Construction workers
City employees	Clerks	Salespeople
Laborers	Professional people	White-collar workers
Barbers	Telephone workers	Others

All persons must be between the ages of 20 and 50. High school and college students cannot be used.

*If you meet these qualifications, fill out the coupon below and mail it now to Professor Stanley Milgram, Department of Psychology, Yale University, New Haven, Conn. You will be notified later of the specific time and place of the study. We reserve the right to decline any application.

*You will be paid \$4.00 (plus 50c carfare) as soon as you arrive at the laboratory.

.....

TO:
PROF. STANLEY MILGRAM, DEPARTMENT OF PSYCHOLOGY,
YALE UNIVERSITY, NEW HAVEN, CONN. I want to take part in this
study of memory and learning. I am between the ages of 20 and 50.
I will be paid \$4.00 (plus 50c carfare) if I participate.

NAME (Please Print)

ADDRESS

TELEPHONE NO. Best time to call you

AGE OCCUPATION SEX

CAN YOU COME:

WEEKDAYS EVENINGS WEEKENDS

Figure 9.1. The ad Milgram ran in the *New Haven Register*. It offered \$4 for an hour of people's time (about \$50 in 2025).

In the decades following Milgram's work, the challenge of finding participants persisted. Over time, researchers came to rely on undergraduate students to participate in studies. But even though undergraduate students solved one problem, they created another: research became limited to primarily young, well-educated people from Western, industrialized nations (Henrich et al., 2010). For the next four decades, up to 80% of studies within fields like experimental psychology were conducted with student participants, raising concerns about how well the findings generalized to people more broadly (e.g., Sears, 1986).

Then, starting in the early 2010's, the development of online recruitment tools dramatically transformed how behavioral scientists find participants (e.g., Buhrmester et al., 2011; Buhrmester et al., 2018; Chandler et al., 2019). Today, most participants are recruited online (e.g., Zhou & Fischbach, 2016). Online platforms give researchers access to millions of people around the world, allowing them to quickly gather data. This change has revolutionized behavioral science in three fundamental ways.

First, online tools have dramatically expanded access to participants. Researchers can easily sample people from different age groups, geographic regions, cultural backgrounds, life experiences, or any other characteristics that are relevant to the research. The diversity of people online solves many of the concerns about the generalizability of findings based primarily on undergraduate student samples (e.g., Mulvaney et al., 2015) and makes it easier to find specific and diverse groups of interest (e.g., Moss et al., 2023).

Second, online methods have made research faster. What took months to accomplish in the past can now be completed in days or even hours. This accelerated pace of data collection allows researchers to test ideas more rapidly, explore new questions, and respond to emerging social phenomena as they happen (e.g., Gharpure et al., 2020; Rosen, 2024).

Third, online tools provide unprecedented flexibility. Researchers can implement a wide range of studies, from simple surveys to complex experiments with multiple waves of data collection. They can track behavior over time, gather data on rare or hard-to-reach populations, and conduct studies that would be logistically impossible in a traditional laboratory, like the example that opened this chapter (e.g., Moss, 2022; Phang et al., 2025).

In some cases, the flexibility of online tools means the research does not have to remain online. In one study, researchers at UCLA, used online platforms to recruit 160 pregnant or post-partum women for an in-person laboratory study with recruitment support from the authors of this book. The research focused on how to build a better bottle nipple for babies with tongue-tie syndrome. Online tools made it far easier to identify people who were eligible and interested in the study than would have been possible otherwise.

Sometimes, the flexibility of online tools allows researchers to mix elements of online research with physical participation. For example, in “I-HUT” studies (in-home usage tests), researchers send participants a specific product, such as a fitness tracker or a device prototype, and collect data about people’s experiences remotely. Alternatively, researchers sometimes ask participants recruited online to send bio-samples, such as saliva, to a laboratory where they can be processed and paired with survey data provided by the participant. Ultimately, whether researchers are conducting purely online studies, laboratory experiments, or some combination of the two, online recruitment extends what is possible in behavioral science. Our goal in this chapter is to explore the options for online participant recruitment and help you understand how best to use each one.

Market Research Panels

When you think about online research, you might picture the kinds of projects from Part I of this book. A student or a professor launches a study, a few hundred to a few thousand participants complete it, and within hours the data are ready to analyze. Although common in academia, these studies represent a small fraction of the research that takes place online.

According to our best estimates, about 5 billion online surveys are completed each year², powering multiple industries and fields of study (Figure 9.2). For example, online surveys drive political polling; they fuel public opinion research on everything from consumer confidence to social attitudes; they help public health researchers understand disease patterns and health behaviors; they enable social science research across disciplines like psychology, sociology, and economics; and perhaps most often, they facilitate market research that guides how companies

² This estimate comes from conversations with the operators of the largest market research panels. By adding the number of surveys that several of the largest panel providers facilitate we arrive at an estimate around 5 billion.

develop products, advertise to consumers, and make businesses decisions. Where does all this data come from?



Figure 9.2. Each year, about 5 billion online surveys are completed around the world, with most of these studies facilitated by market research panels.

Most of it comes from a complex web of businesses and technologies that have been built for market research. You can picture this ecosystem as a three-tiered pyramid (Figure 9.3).

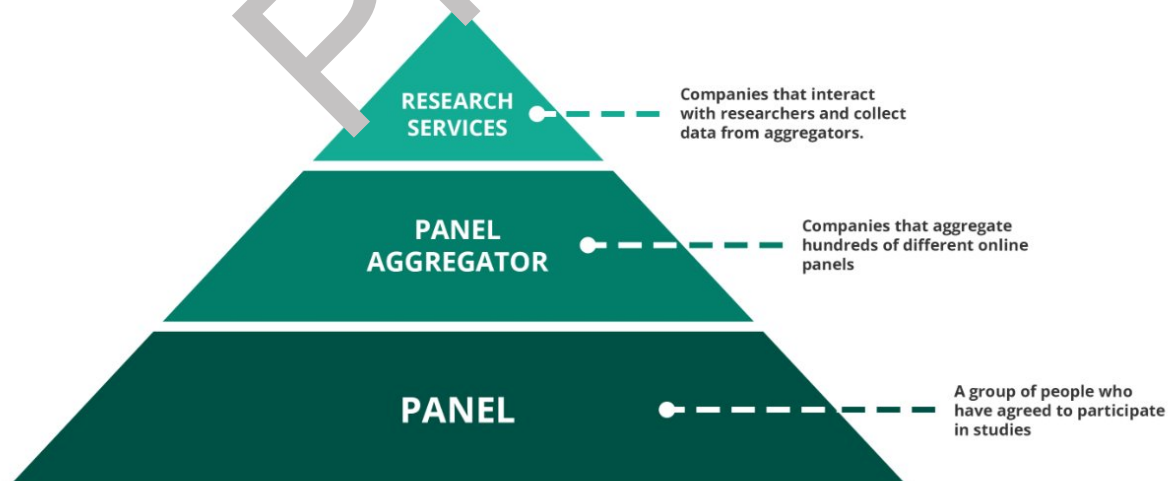


Figure 9.3. The three tiers of the market research ecosystem.

At the base of the pyramid are **panels**. A panel is a group of people who have agreed (opted-in) to be contacted for future research studies. When people sign up with a panel, they give the company their demographic information, contact details, and sometimes their preferences for which kinds of surveys they want to complete. When a research opportunity arises, the panel provider uses its database to invite qualified participants.

Panels range from small operations with a few thousand participants to large companies with millions of members. There are hundreds of panels in the United States and thousands more worldwide. Some of the larger panels have names like Prodege, TapResearch, and Toluna. They host hundreds of millions of surveys every year.

A step above panels, in the middle of the pyramid, are **panel aggregators**. These companies do not maintain their own participants. Instead, they aggregate hundreds of different panels to give researchers greater access to participants. Panel aggregators have developed technology that routes surveys to appropriate panels and integrates the responses from different sources into a single study. This technology is required to find enough people from specific groups to make market research possible. For instance, imagine trying to find 2,000 people who live in Los Angeles, have a household income above \$100,000, and shop for a specific breakfast cereal. It is unlikely that any individual panel would be able to provide these participants, but with aggregation these studies become possible. The most widely used panel aggregator is Lucid, which was acquired by Cint. However, there are other aggregators that are also commonly used. For example, Prime Panels is a panel aggregator by CloudResearch commonly used in academia (e.g. Chandler, et al 2019; Mills et al., 2023).

Finally, at the top of the pyramid are **research services**. Research services sometimes have their own panels or aggregation technology, but far more often they serve as a bridge between researchers and panel providers. When a researcher requests participants from a research service, such as a market research firm, the firm coordinates the request with panel aggregators who pull participants from individual panels. The process looks like Figure 9.4. Within academia, perhaps the most widely known of these services is Qualtrics panels, but many others are available including services from Cint, CloudResearch, and others.

Panel Aggregator in the Online Research Ecosystem

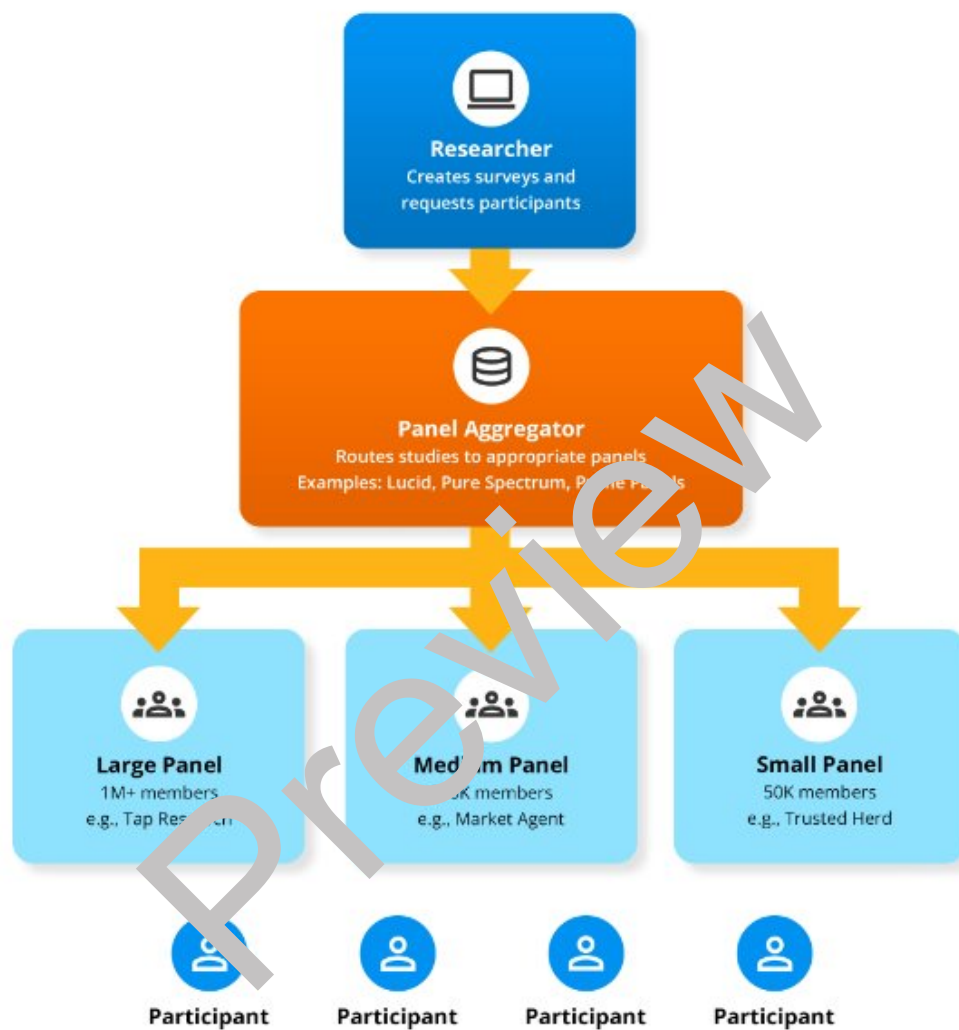


Figure 9.4. How participants are directed to studies within the market research ecosystem.

When researchers use an aggregator, a large sample of several thousand respondents might be drawn from over a hundred different panels, resulting in a supplier list that looks like Table 9.1. Most researchers would likely not recognize the names of these panels and will never learn about their role in the study. But each panel has its own methods for recruiting participants, addressing

issues of data quality, and getting participants to engage in studies. This heterogeneity presents both opportunities and challenges for research.

On the side of opportunity, diversity across panels gives researchers unprecedented access to participants. While one panel might focus on specific groups like those who identify as Hispanic or Latino others focus on geographic regions like China or France. When panels with different specialties are aggregated, researchers can find participants who would not exist in large enough numbers anywhere else, such as African immigrants living in Texas or people with specific medical conditions. They also have greater ability to target people in specific countries, geographic regions, or zip codes.

Another benefit of aggregating panels is that it allows researchers to gather large samples. While most studies require only a few hundred to a few thousand participants, market research panels routinely deliver 10,000 or more people within just a few days. In one study, researchers gathered data from 250,000 people in under two weeks (Fitz et al., 2020).

Prime Insights Group LLC	Consultancy Services LLC	Payswell, LLC	Beijing Youli Technology Co., Ltd.	MindSumo Inc
Make Opinion GmbH	Rewardia.com.au	Reward Holdings	KuRunData	Persona.ly
Qmee	dataspinner	GMO Research	B-old.org, Inc.	Grif & Mif LLC
SocialLoop	Dejour Mobile Research	FusionCash	Splendid Research	Streetbees.com Limited
Publishers Clearing House	Market Cube	IPRoyal FZE LLC	Pick Media Ltd	SurveyEveryone
Branded Research	URWelcome Technologies	Liidimedia oy	Screen Engine/ASI, LLC	A-K International
Bitburst	Verasight	Centiment	Unimrkt Response Inc	Data100
Attapoll	On Device Research	Githaus Research	GAMINGACE TECHNOLOGIES	Ruble Club
Trayistats AI	YouThink.io	Pureprofile	Union Street Enterprises, Inc.	SoftRock
Besitos Corporation, LLC	Bohemian Research LLC	Catalyse Research	Lux Surveys	Quantish Opinion

Unimrkt Research	Innovate	Cre Online	Three Hyphens	Unitedt
Tellwut	Datadvise -	Market Agent	Bridge Money Inc.	Prime Opinions
Dynata US	Research for Good	Hiving	GG2U	SampleBus Market Research Ltd
CashInStyle	99 Ventures	Mo-Web	Idle-Empire	Derrota la Crisis S.L.
Slicethepie ltd	Adnitech LLC	ProductLab	FocusGC	Australian Clearing Pty Ltd
Surveoo-Link	Aeon Research	TapResearch	Rita Personal Data BV	RevenueClick Media PVT Ltd
Rewards1	BizRate	Testable	Rewardingways	FoxyPanel
Drumo	Opivox Panels LLP	Trusted Herd	Samrik Critique Services	Fusion
InboxPounds	Neobux	Kazeel	European Research Tools	Globe Media - MrSurvey
Opinodo	Logit	Gaddin.co	Wash Data Tech Ltd	Hansa Research
Almedia AG	Mypinio GmbH	Insight works	XP Interactive, LLC	Ignite Vision
Dale Network	Competition Panel	Survey Promo	Nomadic Insights, LLC	InBrain.ai
Opinaia Panel	Edison Meti	Syno Rewards	TGM Research FZE	Inginit PTE LTD
Alpha Poll	Shopbook	Apex	The Coupon App, LLC	Savvy Technology - Match
Purely Research	iSurveyPanel	Bovitz, Inc.	Research on Mobile	Shanghai Wanyan Network
Aspen Analytics	Poll Pronto	Daily Rewards	A One Market Research	MadCashSurvey - Match
ITC	Maiwen China	Walnut Unlimited	Madai - API	Maholla
TheoremReach	iAngelic	Surveyeah	MySoapBox	Mewug GmbH
MDQ	GrowthOps	Promio	Offernation	MindMover

Table 9.1. Panels that supply participants to market research aggregators.

The size of market research panels is one reason they are used so often for polling. For example, leading up to the 2024 U.S. Presidential election, the Siena College Research Institute (the organization that does polling with *The New York Times*) collaborated with the authors of this book to collect data through PrimePanels, an aggregator offered by CloudResearch (Chandler et al., 2019; Moss et al., 2023). The researchers surveyed 6,000 participants across seven battleground states, with each state's sample matched to demographic and county distributions of the state (Siena College Research Institute, 2024). This kind of targeting would not be possible without aggregation.

Yet, market research panels have disadvantages, too. The most serious disadvantage is data quality. In a typical study where participants are sourced through market research panels or panel aggregators, over 40% of participants can be expected to provide unusable data (e.g., Litman et al., 2023; Stagnaro et al., 2024, Weber, 2023). We discuss the quality of online panels in Chapters 10 and 11, but the lack of standardization in how panels recruit and vet participants means that if researchers do not take steps to address data quality within their studies the results can be unusable, misleading, or outright false (e.g., Gharpure et al., 2020; c.f. Litman et al., 2023).

Another challenge of market research panels is that researchers cannot control how much participants are paid, and overall compensation tends to be low. In a typical study, participants will receive something equivalent to perhaps ten to fifty cents for their time. While this model works for short and simple surveys, it limits how much effort people are willing to invest in longer and more difficult studies.

Finally, a third challenge of market research panels is that researchers cannot directly communicate with participants. All interactions are mediated through the panel provider, meaning researchers cannot give participants special instructions or build the rapport necessary for complicated projects. Participants from market research panels are unlikely to complete studies that require long open-ended responses or complex tasks like downloading and using apps. They are also unlikely to stick with longitudinal studies. About 60% of participants will

drop out, or attrit, in the first week of a longitudinal project) and that number increases to over 70% in a 30-day period (see Chapter 14 for more detail). The type of study we described at the start of this chapter, where 1,000 participants completed daily tasks for 30 days, would be impossible through the typical market research panel.

Due to these limitations, most of the academic research conducted online uses more specialized panels that we refer to as researcher-centric online platforms. We will explore these platforms in the next section. Even though researcher-centric online platforms are a better fit for most academic studies compared to market research panels, it is important to understand the market research ecosystem for several reasons.

First, given how often these panels are used, you will encounter data from these sources in the future whether in published studies, news reports, or policy documents. Second, you may use these panels to reach niche samples that cannot be found elsewhere. Finally, knowing about this ecosystem, and its strengths and weaknesses, will help you critically evaluate research findings you encounter.

The last thing to note about market research panels is that many companies operate across multiple tiers of the ecosystem presented in Figure 9.3. For example, CloudResearch offers both its own panel platform, Connect, and an aggregator service called PrimePanels. These products serve different research needs. Whereas Connect is a researcher-centric online platform which provides direct access to a carefully vetted pool of high-quality participants, PrimePanels offers access to specialized demographic groups, international sampling, and large sample sizes.

More generally, cross-tier integration is increasingly common because most panels do not operate in isolation but collaborate with other panels. When a market research panel receives a request for participants with characteristics that are underrepresented in their own pool, they might partner with another panel to fulfill the request rather than turn the study down. This interconnectedness helps explain why the boundaries between different parts of the ecosystem often appear blurry. Researchers might work with what they believe is a single panel, unaware that their participants are sourced from multiple providers behind the scenes. By understanding the fundamentals of this ecosystem, you will be better equipped to make informed decisions

about your own research. Table 9.2 summarizes the strengths and weaknesses of market research panels and panel aggregators.

Strengths	Weaknesses
Scale and Reach Access to tens of millions of participants globally, making it possible to collect samples of 10,000+ respondents in days.	Data Quality 40% or more respondents may be low quality or fraudulent without proper quality control measures.
Geographic Targeting Ability to collect data from different geographic areas, down to the zip code level.	Limited Engagement Low compensation results in participants who give minimal effort, especially in complex tasks.
Demographic Specificity Possible to find niche populations thanks to aggregation across hundreds of panels.	Mediated Communication Researchers cannot directly communicate with participants, limiting the ability to provide special instructions or clarifications.
Census Matching Sufficient volume to match samples to census demographics in specific states.	Time Constraints Studies work best when 20 minutes or less.
Speed Rapid data collection, with very large and niche population studies often completed quickly.	High Attrition Around 60% of participants drop out in the first week of longitudinal studies.
Cost Efficiency Relatively affordable compared to traditional research methods like telephone surveys or in-person interviews.	Limited Response Types Not suited for long open-ended responses or complex interactive tasks.
Quota Controls Sophisticated tools for setting and managing quotas across demographic characteristics.	Limited Flexibility Researchers can't adjust compensation or incentives to improve participation.
Accessibility Available to researchers without specialized recruitment expertise or established participant pools.	Technical Limitations Hard to implement studies that require downloads, apps, or specialized software.
Research Diversity Suitable for a wide range of descriptive and correlational research across many disciplines and industries.	Longitudinal Challenges Nearly impossible to conduct extended studies with daily participation over weeks or months.

Table 9.2. The strengths and weaknesses of market research panels.

Researcher-Centric Online Platforms

While there are hundreds of market research panels, most academic research occurs on just a few platforms, such as CloudResearch Connect, Amazon Mechanical Turk (MTurk), or Prolific.

These platforms are best thought of as **researcher-centric platforms** because they allow researchers to control the research process to a greater degree than market research panels.

The first researcher-centric platform to be widely used for behavioral research was Amazon's Mechanical Turk (MTurk; Buhrmester et al., 2011; Paolacci et al., 2010). Although it was not designed specifically for research, MTurk single-handedly moved academic research online (Litman & Robinson, 2020; Moss et al., 2024). Then, beginning in 2018, the quality of data on MTurk declined substantially, causing behavioral scientists to look for new sources of participants (Hauser et al., 2023; Moss & Litman, 2018; see Chapter 10). Some of the alternatives they found, like Connect and Prolific, were specifically built for behavioral research (O'Grady, 2024).

At first glance, researcher-centric platforms appear similar to market research panels, as both maintain a database of participants (a panel) willing to complete tasks for compensation. However, unlike market research panels, researcher-centric platforms put the researcher in full control of the study.

The first factor that differentiates researcher-centered platforms from market research panels is the ability to specify participant compensation. Controlling compensation allows researchers to conduct more complicated projects because when people feel they are fairly compensated for their time they are more willing to complete long, demanding, or difficult tasks (e.g., Litman & Robinson, 2020; Moss et al., 2023).

For example, Figure 9.5 shows where researchers set compensation on Connect. Because the participants are most often motivated by financial rewards, researchers can find people willing to participate in video interviews, test new websites, analyze text, work in teams, discuss political issues, review scientific papers, suggest research ideas, track their experiences over time, and much more, as long as they are well compensated (Arechar et al., 2018; Boynton & Richman, 2014; Campbell & Reiman, 2022; Gallo & Gran-Ruaz, 2021; Garbinsky et al., 2020; Hall et al., 2020; Kittur et al., 2008).

Figure 9.5. Researchers using Connect decide how much they want to pay each participant. This feature helps motivate people to complete difficult or demanding projects.

Second, researcher-centric platforms give researchers the ability to communicate with participants. Often, this communication occurs through a messaging or email system like the “Conversations” page on Connect (Figure 9.6).

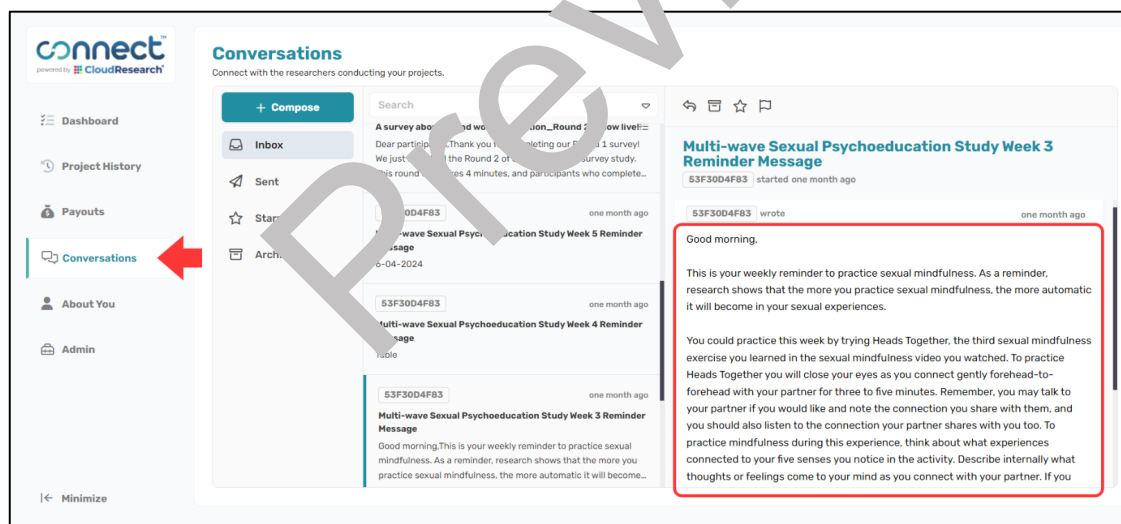


Figure 9.6. The Conversations center on Connect allows researchers to send email messages to participants while maintaining participant anonymity.

Communication with participants can enhance the quality of a study. Most tangibly, communication allows participants to tell researchers when something is wrong such as a broken

link, a question that does not make sense, or an issue that prevents participants from doing what the researcher has asked. Communication also allows researchers to provide participants with support, and it can be part of upholding the ethical standards of behavioral research discussed in Chapter 15.

Even so, the most common reason to communicate with participants is to remind them about subsequent rounds of data collection in a longitudinal study. For instance, the message in Figure 9.6 is from a longitudinal study that asked participants to engage in a sexual mindfulness intervention. The text of the message reminds participants to practice sexual mindfulness—being mentally present and nonjudgmental—during intimate moments with their partner, such as while cuddling or holding hands. Reminders such as these increase participant engagement and improve retention rates in longitudinal studies, especially when paired with the ability to control how much participants are compensated (e.g., Hall et al., 2020).

Finally, crowdsourcing sites offer significantly higher data quality than other recruitment options (e.g. Stagnaro et al., 2024), as discussed in more detail in Chapters 10-12. The main reason for this is that researcher-oriented platforms vet participants before they can participate in studies. If you created a Connect account in Chapter 9, you experienced this vetting. Research platforms that vet participants produce better data quality than those that do not (e.g., Peer et al., 2023; Stagnaro et al., 2024).

Together, the strengths of crowdsourcing sites—control of compensation, the ability to communicate with participants, and high data quality—allow researchers to conduct a variety of projects that go beyond simple surveys. These include interactive experiments and studies where participants need to download software for cognitive research that records reaction times with millisecond precision (e.g., Stewart et al., 2017). It also includes video interviews, real-time interactions between participants (e.g., Keller et al., 2023), mock jury trials (e.g., Salerno et al., 2023), complex tasks like content analysis of written texts (e.g., Benoit et al., 2016), and multi-session longitudinal studies (e.g., Hall et al., 2020). In fact, longitudinal studies are so common that Connect offers a feature called Waves for managing these projects (see Chapter 14).

Researchers can use Waves to create studies with dozens or even hundreds of follow-up points

and automate the work of launching studies and reminding participants to participate. Features like this elicit unprecedented engagement from participants.

The study that opened this chapter is a good example of what differentiates a researcher-centered platform like Connect from a typical market research panel. Researchers had to interact with participants every day, choose how much to pay people, and create a workflow for tracking who participated in previous studies. The researcher dashboard on Connect helped facilitate all these tasks.

The main limitation of researcher-oriented platforms is that they cannot provide access to as many niche groups as market research panels (e.g., Moss et al., 2023). This is a direct result of aggregation. Because researcher-oriented platforms operate as a single panel, there are many groups of participants researchers simply cannot reach or may not reach in the numbers they desire. Often, the biggest limitations arise when sampling small groups in the population and people from specific, harder to reach geographic areas such as sparsely populated states or communities.

Despite some limitations on who can be reached, the overall combination of high data quality, researcher control, methodological flexibility, and participant engagement explains why academic researchers have embraced researcher-oriented platforms so enthusiastically. For studies requiring thoughtful responses, complex designs, or longitudinal follow up, the advantages often outweigh smaller participant pools compared to traditional market research panels. Table 9.3 summarizes the strengths and weaknesses of researcher-oriented platforms.

Strengths	Weaknesses
<p>Superior Data Quality Rigorous verification procedures reduce fraudulent responses, compared to up to 40% in traditional market research panels (e.g., Stagnaro et al., 2024; Moss et al., 2023).</p>	<p>Smaller Participant Pools Fewer total participants than major market research panels, which limits sample sizes for very specialized populations.</p>
<p>Researcher Control Researchers can set compensation rates, implement custom screening, and directly communicate with participants.</p>	<p>Potential Non-naivety Some participants may have more research experience, affecting the results of some study designs.</p>

Strengths	Weaknesses
<p>Methodological Flexibility Supports complex study designs including longitudinal research, interactive experiments, specialized tasks beyond simple surveys, studies requiring downloads of specialized software, video interviews, or interactive studies.</p>	<p>Generally Less Representative Samples Demographics may not naturally match census proportions without implementing specific quotas.</p>
<p>Participant Engagement Participants are more attentive, thoughtful, and willing to engage with complex or time-consuming tasks compared to other online sources.</p>	<p>Limited Global Coverage Most sites are limited to just a few countries or a single region</p>
<p>Longitudinal Capabilities Higher retention rates for multi-session studies, makes extended research designs more feasible.</p>	

Table 9.3. The strengths and weaknesses of crowdsourcing platforms for research.

Representativeness in Online Sampling

Module 9.2

Examine issues of sampling and how to make online samples more representative

In the previous section, we described the diverse sources for recruiting participants online. The present module delves into the concepts of sampling and representativeness as they relate to both online and offline samples. We will begin by defining what constitutes a representative sample and why it matters for research. Then, we will examine the two primary approaches to sampling participants—probability and non-probability—and discuss why most behavioral research uses non-probability sampling. A key focus will be on understanding common practices like quota sampling, its utility in online studies, and the important distinctions between these demographically adjusted samples and true probability samples. Finally, we will explore the "fit-for-purpose" framework, which is an essential guide for thinking about sample representativeness.

What is a Representative Sample?

Most studies in the behavioral sciences examine a **sample**, which is a specific group of people selected from a larger **population**. A **population** is the group of people a researcher wishes to understand and draw conclusions about. For instance, if a researcher was interested in the political attitudes of adults in the United States, the population would be all adults in the U.S. The researcher might then survey a sample of a few thousand people.

A population does not have to be the size of an entire country. Instead, students at a specific university or people who drink Coca Cola can be considered a population, so long as those are the people the research is focused on. The critical question for any study with a population of interest then becomes: how well do the people in the sample reflect the population? This is the essence of **representativeness**. A truly representative sample accurately mirrors the characteristics of the population it is drawn from.

Two Approaches to Sampling: Probability and Non-Probability

Researchers generally use one of two broad strategies for selecting participants: probability sampling or non-probability sampling.

Probability Sampling

Probability sampling is designed to create a sample that is a statistical mirror of the larger population. Its defining feature is that every person in the target population has a known, and roughly equal chance of being selected to participate. Think of it like a lottery where every member of the population holds a ticket.

To collect a probability sample, researchers typically start with a nearly complete list of everyone in the target population, known as a **sampling frame**. A sampling frame might consist of a list of phone numbers of people in the country. Participants are then randomly selected from this sampling frame. If this process is executed correctly, the resulting sample should naturally reflect the distribution of characteristics found in the overall population. For instance, if we were studying social media use in the United States where usage varies from very little to very frequent, a true probability sample would capture heavy users, light users, and everyone in between in proportions similar to their actual presence in the U.S. population.

As with any statistical process, there is some error associated with measuring people's attitudes or behaviors. In probability sampling, this is referred to as the **margin of error**, or, how much the results of the survey can be expected to vary by chance. The margin of error in probability samples is a function of the sample size; the larger the sample, the smaller the margin of error. For example, in a sample of 500 people the margin of error is approximately 4%, in a sample of 1,000 it is 3% and in a sample of 2,000 it is 2%.

Putting all of that together, if a random sample of 1,000 people drawn from the US population showed that a President's approval rating is 40%, it would mean that 40% of people in the United States plus or minus 3% (somewhere between 37-43%) approve of the President's performance.

Non-Probability Sampling

In contrast to probability sampling, non-probability sampling is a way of selecting participants where not everyone in the population has an equal or known chance of being included. Instead, people might be included because they are readily available, referred to as a **convenience sample**. Traditionally, most research in behavioral science has used undergraduate participants in introductory psychology and other classes. These students are a convenience sample.

Online panels are also a source of non-probability samples because participants are self-selected rather than chosen through a random process from the population. People who self-select into panels are generally different demographically from those who do not. For example, they tend to be more technologically knowledgeable, more educated, more likely to be female, and more likely to lean left politically than the general population (see Litman and Robinson, 2020).

Somewhere in-between probability samples and pure convenience samples are **quota-based samples**, sometimes referred to as **purposive samples** (see Baker, 2013). In quota sampling, researchers aim to match their sample to the population based on known attributes, such as age, gender, education level, or geographic region.

Quota sampling is a standard practice in much of online research, where researchers will set quotas to align their sample demographics with those of the U.S. Census. For instance, if the Census indicates that 51% of the population is female and 13% is African American, a researcher using quota sampling would aim to recruit participants until approximately 51% of their sample is female, 13% is African American and so on (Chapter 14 demonstrates how this is done). Thus, while the online panel is itself selected from a non-random pool of people, the sample demographics will match the population of the United States.

While quota sampling can make a sample look more like the population on the selected demographics, it does not transform a non-probability sample into a true probability sample. The fundamental difference lies in the initial selection process. Probability sampling gives every person in the sampling frame a known and equal chance of being selected *before* quotas are considered. Quota sampling, on the other hand, selects from an already non-random pool of people and then matches the final proportions to U.S. Census targets.

The key technical characteristic of non-probability samples is that there is no known or agreed upon way to measure the margin of error (see Baker, 2013). Regardless of where the participants come from, there is no way to know how much the sampling frame for a non-probability sample differs from the population. For this reason, it is significantly more difficult to make inferences about the frequency with which something occurs in the population from a non-probability sample than a probability sample.

Why Behavioral Research Often Relies on Non-Probability Samples

Despite the strengths of probability sampling, most studies in the social and behavioral sciences, both online and offline, use non-probability samples. Think about the two studies that opened this chapter: both the classic Milgram study on obedience to authority, and the longitudinal study by open AI. Neither of these studies used probability samples.

Neither have most studies throughout the history of the behavioral sciences. In 1986, David Sears reported that approximately 85% of studies published in social psychology journals used undergraduate students and two decades later a similar note was found in consumer research (Peterson, 2001). There are three reasons for the long history of using non-probability samples: 1) in experimental research, internal validity is often prioritized over external validity; 2) in the vast majority of cases, the experimental effects and associations observed with probability samples are replicated with non-probability samples, and 3) non-probability samples allow for more complex research designs and better access to hard-to-reach groups.

Testing Treatment Effects

Most behavioral science studies are either experiments, such as those described in Chapter 7, or investigations of the relationships between variables, such as those described in Chapters 5 and 6.

In experimental studies, researchers often prioritize internal validity and the ability to detect whether an effect is present over obtaining a perfectly representative sample (see Hayes, 2017). This is because random assignment of participants to different experimental conditions nullifies any pre-existing differences across groups. In other words, even if the participants do not mirror the larger U.S. population, a well-executed study can still establish a cause-and-effect relationship.

Indeed, an experimental effect observed on a non-probability sample is often just as informative as one observed on a probability sample. Consider, for instance, the process of testing a new drug in a clinical trial. The primary goal is to determine if the drug works. That is, does the drug have the intended therapeutic effect compared to a placebo or other existing treatments? In such experimental studies, the most critical factor for establishing causality is random assignment of participants to the treatment and control conditions. When participants are randomly assigned, pre-existing differences between the groups, whether demographic, medical history, or other unmeasured variables, are distributed equally across the treatment conditions. This allows the researchers to attribute any observed differences in outcomes to the drug itself rather than to other factors.

For this reason, initial drug trials are often conducted on samples that do not represent the entire population that might eventually use the drug. Instead, most drug trials rely on participants who are interested, available, and reside in a convenient geographic region that is close to the physical location of the trial. The same logic that works for drug trials works for psychological and behavioral experiments. Thus, many behavioral studies use convenience samples because doing so does not impede their ability to draw meaningful conclusions.

Comparisons of Treatment Effects on Probability and Non-probability Samples

In addition to the importance of internal validity, a growing body of research shows that online convenience samples and probability-based, representative samples yield similar results across the vast majority of studies.

In recent years, several large-scale studies have systematically compared experimental treatment effects found in online non-probability samples with those from nationally representative probability-based samples, such as the Time-Sharing Experiments for the Social Sciences (TESS, 2024). In one study, for instance, researchers examined 23 experiments from across the social sciences and found that more than 80% of the 36 treatment effects they examined on Mechanical Turk perfectly replicated those from probability-based samples (Mullinix et al., 2015). Importantly, in none of the experiments did the MTurk results show effects in the opposite direction of the probability-based samples, suggesting a low risk of **Type I** (finding an effect that doesn't exist) and **Type II** (not finding an effect that does exist) errors when using online convenience samples.

Similar research by Coppock et al. (2018) compared 27 experimental treatment effects across hundreds of demographic subgroups in MTurk and probability samples and found that the results were in the same direction in virtually all comparisons, with only a small percentage differing significantly in magnitude. Similar consistency of results has also been found in correlational studies. In one instance, researchers compared the associations between various political attitudes (such as egalitarianism, moral traditionalism, and authoritarianism) in MTurk samples to those found in data from the nationally representative American National Election Studies (ANES; Clifford et al., (2015). Across 72 different effect sizes, the researchers found 68 were similar in both samples, with the few statistically different effects still showing relationships in the same direction. Thus, the overall evidence from these studies suggests that for most experimental and associative research, online convenience samples are likely to yield conclusions similar to those derived from probability samples.

Complex Designs and Better Access to Niche Groups

The final reason non-probability samples are common in the behavioral sciences is because they allow researchers to conduct the complex projects described in Module 9.1 or to reach niche groups. When researchers want to test ideas or study sub-groups in the population, it is often easier to find people from the market research panels or researcher-centric platforms we discussed earlier than it is to use a form of probability sampling. In other words, participants from non-probability sources are often a better fit for the kinds of studies most behavioral scientists want to conduct.

The Fit-for-Purpose Framework: Matching Sample to Research Goal

Given all the considerations of sampling participants, how should researchers think about probability and non-probability samples. The **fit-for-purpose framework** (Baker, 2013) provides a useful and practical approach. This framework suggests that rather than viewing samples as simply "good" or "bad," "representative" or "unrepresentative", researchers should think about how well a sample meets the goals and design of a research project. Under this framework, different questions require different samples.

According to the fit-for-purpose framework, the choice of participant source and sampling strategy should be driven by the specific research question and objectives. Online non-

probability samples generally demonstrate a high fit-for-purpose for experimental and associative research. When the primary goal is to understand the nature and strength of a relationship, nonprobability samples offer a good ‘fit-for-purpose’ approach. Indeed, as described above, comparative studies have shown encouraging consistency in the effects found in non-probability panels versus those in nationally representative samples.

On the other hand, for studies that aim to provide a precise population estimate, such as political polls, non-probability online panels are generally not the best fit. Because these samples are opt-in and may have demographic compositions different from the general population, direct frequency estimates can be biased.

Nevertheless, in practice, such non-probability online panels are commonly used in polls that aim to precisely measure the general population. Indeed, most polling is currently conducted with non-probability online panels (see Cohn, 2024; Kennedy, 2013; Kennedy et al., 2020). In such cases, **quota sampling** is combined with **statistical weighting** after data collection, which can substantially improve the accuracy of estimates from non-probability samples. It is generally agreed, however, that such methods may not remove all bias (see Baker, 2013; Tourangeau et al., 2013; Stagnaro, 2024).

Complex Research Designs

Finally, the discussion of sampling methods and representativeness must also consider the practical feasibility of conducting different types of research. Recall the study described at the beginning of the last chapter, where OpenAI and the MIT Media Lab investigated the psychological effects of frequent interactions with ChatGPT. That research involved nearly 1,000 participants engaging in daily tasks and interactions with ChatGPT over a one-month period, with researchers measuring outcomes like loneliness and emotional dependence on AI.

Consider the immense practical challenges of conducting such an intensive longitudinal study using traditional probability sampling methods aimed at achieving a perfectly representative national sample. Recruiting nearly a thousand people through probability sampling, and then following their daily engagement with specific tasks and AI interactions for an entire month, would be extraordinarily complex and prohibitively expensive.

This is where online convenience samples demonstrate their unique value. These platforms are specifically designed to facilitate complex research studies. They allow researchers to recruit large numbers of participants relatively quickly, manage communication for longitudinal studies, implement systems for daily reminders or task delivery, and incentivize sustained participation in demanding or lengthy projects. Without such infrastructure, research questions requiring intensive longitudinal data collection or complex interactive designs might remain unexplored due to practical and financial barriers.

Summary

Behavioral scientists can access online research participants through multiple channels, each with distinct advantages. Researcher-centered platforms offer versatile tools for various research designs, strong quality controls, and participants willing to engage in these tasks. Market research panels excel at complex demographic targeting and large samples.

When the primary goal of research is to test predictions from theories—whether through experiments or the examination of associations—online non-probability samples have demonstrated a strong fit-for-purpose. The findings regarding experimental effects and correlations are reliable and have been accepted as standard practice across many disciplines in behavioral science, particularly when researchers employ strategies like quota sampling where appropriate.

The most effective sampling approach depends on your specific research needs. When selecting a participant source, the fit-for-purpose framework is a helpful and practical approach when thinking about which source to use

Additional Readings

- Buhrmester, M. D., Talaifar, S., & Gosling, S. D. (2018). An evaluation of Amazon's Mechanical Turk, its rapid rise, and its effective use. *Perspectives on psychological science, 13*(2), 149-154.
- Chandler, J., Rosenzweig, C., Moss, A. J., Robinson, J., & Litman, L. (2019). Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior research methods, 51*, 2022-2038.
- Coppock, A., & McClellan, O. A. (2019). Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Research & politics, 6*(1), 2053168018822174.
- Hartman, R., Moss, A. J., Jaffe, S. N., Rosenzweig, C., Litman, L., & Robinson, J. (2023). Introducing Connect by CloudResearch: Advancing online participant recruitment in the digital age.
- Moss, A. J., Hauser, D. J., Rosenzweig, C., Jaffe, S. N., Robinson, J., & Litman, L. (2023). Using market-research panels for behavioral science: An overview and tutorial. *Advances in Methods and Practices in Psychological Science, 6*(2), 25152459221140388.

Chapter 10: Data Quality

The National Longitudinal Study of Adolescent to Adult Health, abbreviated Add Health, is the largest, most comprehensive survey of its kind ever conducted. Funded by the U.S. government, it began during the 1994-95 school year with a nationally representative sample of over 20,000 students in grades 7-12, and it has continued with five additional waves of data collection to date.

Over the years, Add Health has gathered a stunning amount of information about participants' demographic, social, familial, socioeconomic, behavioral, psychosocial, cognitive, and health characteristics, as well as that of their parents. To this self-reported data, researchers have added information about participants' schools, neighborhoods, and communities. During in-home visits, researchers have gathered physical and biological data, including genetic markers, blood-based assays, body measurements, and information about people's medications. Simply stated, Add Health is a remarkable achievement of behavioral research that has required hundreds of millions of dollars and the effort of tens of thousands of people. In terms of methods and rigor, it's as good as it gets.

Yet, it is precisely because of that rigor that researchers were surprised to find people who lied about their data. On several occasions, researchers found significant discrepancies between what participants reported in survey and what was evident during follow-up visits at participants' homes. In one striking example, researchers identified 253 participants who claimed to use artificial limbs. But, when the researchers visited these people in person, only two had such limbs. The other 99% had lied (Fan et al., 2006).

Unfortunately, this was not the only problem with the data. In another study based on Add Health data, researchers found that adopted children were at higher risk than non-adopted children for failing school, using drugs, getting into fights, lying to their parents, and experiencing physical and mental health problems (Miller et al., 2000). During subsequent in-home visits, the researchers discovered that approximately 20% of the students who reported being adopted had lied; they were living with their biological parents (e.g., Fan et al., 2006). When these "mischievous participants" were removed from the dataset, the differences between

adopted and nonadopted adolescents disappeared. Similar problems were found for self-reported immigration status and other outcomes (see Fan et al., 2006)

These problems with data quality are not unique to the Add Health database. In fact, they are as old as survey research itself. In the 1970s, for instance, researchers noticed a small but consistent number of people who claimed in surveys to take drugs that did not exist (e.g., Pape & Storvoll, 2006; Petzell et al., 1973). In other studies, researchers have found participants who lied about their sexual orientation, their gender or age, experiences with consumer products, gang affiliation, whether they have a vision impairment, whether they have ever been pregnant, and whether they own pets, among many other things (e.g., Chandler & Paolacci, 2017; Hartman et al., 2023; Robinson-Cimpian, 2014; Wessling et al., 2017). In each instance, removing unreliable participants changed the study results.

While problems with data quality exist in all research environments, they pose an especially large problem in online studies where up to 40% of data comes from problematic or unreliable sources. In this chapter, we explore the scope of data quality problems in online research. In Module 10.1 we look at how poor-quality data can produce misleading claims that spread through both academic and public discourse. Then, we dive into the global network of online survey fraud, showing you how click farms and other scams operate to take the money offered for participating in studies.

As part of this conversation, we tell the story of Mechanical Turk. Problems with fraud on MTurk became particularly evident in 2018 when researchers experienced a sudden drop in data quality that threatened research across multiple disciplines. This crisis highlighted how vulnerable online research platforms can be to systematic fraud. In response, specialized platforms have emerged to replace MTurk with rigorous verification procedures and more sophisticated detection systems. These platforms typically reduce fraudulent data to less than 5% of responses, compared to the 30-40% found on MTurk and typical market research panels.

After describing problems with data quality, Module 10.2 examines how these problems affect descriptive, correlational, and experimental research. Through real-world examples, you will see how poor data quality can inflate point estimates in descriptive studies, create false correlations in associative research, and dilute effect sizes in experimental work. We will examine how these

distortions can lead even careful researchers to draw incorrect conclusions about human behavior.

By understanding both the nature of data quality problems and their consequences for different types of research, you will be better prepared for the next chapter, where we introduce practical solutions for detecting and addressing these issues. The goal is to enable you to conduct online studies that yield reliable, meaningful insights into human behavior on any online platform.

Preview

Chapter 10 Outline: Data Quality

Module 10.1: Data Quality in Online Research



Reading

Examine the threats to online data quality

- Explore several high-profile research mistakes that were due to data quality
- Learn about the global network of survey fraud
- Understand the rise and fall of data quality on Mechanical Turk

Module 10.2: How Data Quality Affects Research



Reading

- **Learn how low-quality data distorts descriptive, correlational, and experimental research**
- Examine how poor data distorts descriptive research
- Look at what poor data does to correlational studies
- Examine how poor data wrecks experimental research

Data Quality in Online Research

Module 10.1

Examine what the threats are to online data quality

In recent years, public polling has revealed the following things about the U.S. population:

- 20% of people support political violence (National Public Radio, 2023)
- 30% of Millennials are unsure if the Earth is round (Scientific American, Nguyen, 2018)
- 20% of Millennials think the Holocaust is a myth (The Economist/YouGov, 2023)
- 4% of people drank or gargled bleach to protect against COVID-19 (Centers for Disease Control, CNN, 2020)
- Over 50% of African Americans don't think it's okay to be white (Rasmussen, 2023)

These results are unflattering. They suggest people in the U.S. deny basic science, embrace conspiracy theories, support political violence, are ignorant of history, and are divided by race.

But there is good news: all these findings are false.

Each finding above stems from data quality problems that went undetected. Because the researchers failed to detect the problems, not only were they misled about what people think, but so were the scientific journals and news outlets that published the findings and the citizens who consumed them.

The impact of misleading findings extends far beyond academic circles. Misleading research can shape public opinion and influence policy decisions. For instance, take the claim about Millennials doubting the Earth is round. This finding appeared in *Scientific American*, one of the most trusted popular science publications in the world. Once there, it gained tremendous credibility. It was later mentioned several times on the popular podcast, *The Joe Rogan Experience*, where Joe Rogan discussed the finding with renowned astrophysicist Neil deGrasse Tyson (Figure 10.1). This exchange, viewed by tens of millions of people, has the potential to legitimize a false narrative about how young Americans view the world.

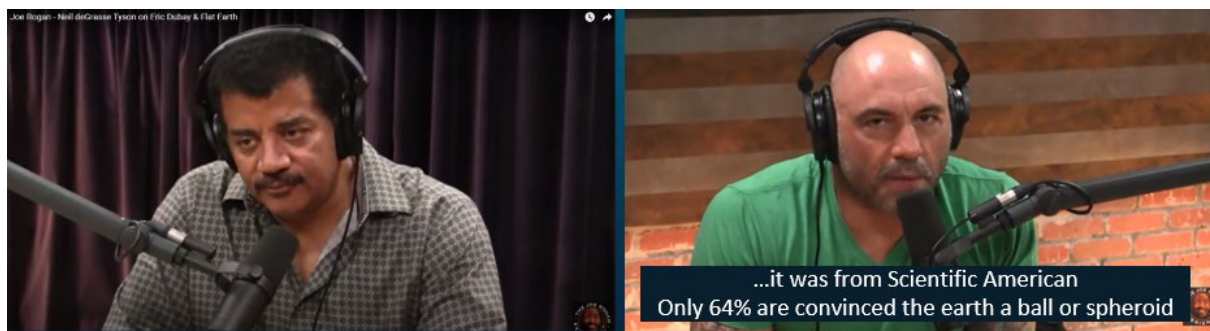


Figure 10.1. A false finding about people’s doubt in the shape of the Earth spread so far that Joe Rogan asked Neil deGrasse Tyson about it on an episode of *The Joe Rogan Experience*.

A similar pattern has played out with other findings. The claim about Millennials denying the Holocaust prompted discussions about education policy and historical awareness. The statistics about Americans consuming bleach raised concerns about dangerous health practices during a global pandemic. And the finding about racial attitudes fueled already tense conversations about race relations in America, even though the real issue was one of measurement. How did these things happen?

The answer lies in data quality. Remember that about 5 billion surveys are completed online each year, and the majority occur through market research panels. This massive industry has transformed how scientists gather data about human behavior while also creating unprecedented challenges for data quality.

The best estimates today are that around 40% of responses in a typical online study come from unreliable sources (Chandrasekhar et al., 2019; Litman, et al., 2023; Mercer et al., 2024; Weber, 2023), although the number can sometimes be much higher (Stagnaro et al., 2024).

In one study, an industry group formed to examine data quality, called Case for Quality, collected data from over 4,000 respondents from four different online panels. They found that between 30 and 40 percent of data could not be used for analysis, either because of fraud or inattention. Fraud was equally apparent across all panels and took numerous forms. One notable behavior was that respondents made implausible claims in the survey. For example, the study was conducted during the height of the COVID-19 pandemic in the United States, when all major theaters and entertainment events were closed. Yet, many people in the survey claimed to have gone to the opera in the last month.

Similarly implausible claims have been found in other studies. For instance, when researchers examined data quality across three commonly used online panels with a sample of 2,500 respondents, they included a question about whether people had recently purchased a home in McMullen, Alabama. According to the 2020 Census, McMullen has a population of thirty-two people. Nevertheless, 417 participants reported recently purchasing a home there, clearly indicating they were not reading the questions or intentionally lying (Reavey et al., 2024).

There has been considerable speculation about who or what lies behind unreliable online data. The most common explanation has been that problematic data comes from ‘bots,’ automated computer programs designed to fraudulently take online surveys. Despite this widespread belief, research suggests most problems with data quality come from people (e.g. see Jaffe et al., 2025; Kennedy et al., 2020).

Direct evidence for human respondents comes from several sources. In one case, researchers identified participants whose past responses suggested fraud and invited them to a Zoom interview (Jaffe et al., 2025). When these participants joined the video calls, what researchers saw was revealing. As shown in Figure 10.2, the participants appeared in rooms with multiple computers where other people were taking surveys. The people in the room spoke languages other than English and were outside of the United States despite participating in studies restricted to the U.S.

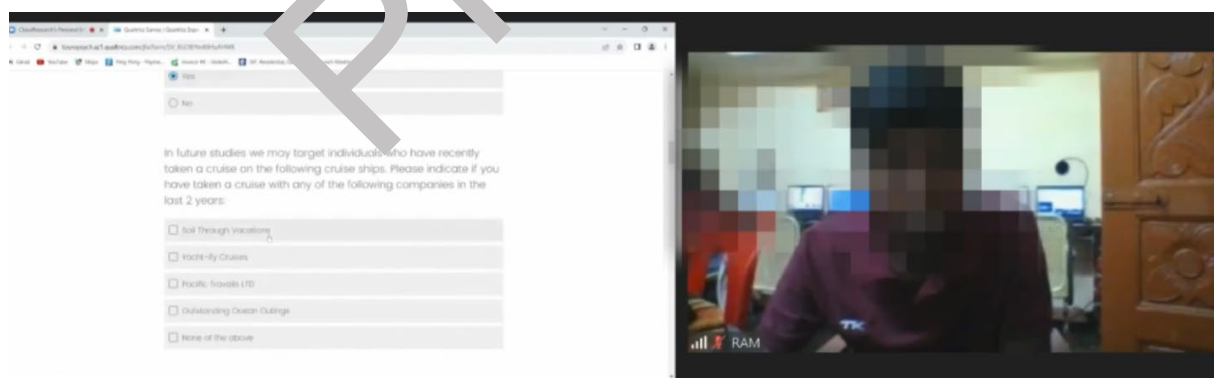


Figure 10.2. A look inside a survey click farm.

Particularly revealing was how participants in these interviews answered survey questions. During the interviews, the researchers opened a Qualtrics survey and asked participants to

complete the survey while sharing their screen. Even though the participants knew they were being observed, they provided implausible answers.

For example, when asked if they had recently purchased a home in McMullen, Alabama—the town with just 32 residents—they answered "yes." When presented with a list of cruise lines and asked which ones they had traveled on recently, the participants selected fictional cruise lines that do not exist. They also reported using products and services that do not exist and said they had experienced several unlikely events such as filing recently for homeowners' insurance due to a lightning strike.

One reason the participants in these interviews gave such unlikely answers is because facilities like those in Figure 10.2 are what is commonly called a "click farm." Click farms operate all over the world, in countries like India, Bangladesh, Russia, Nigeria, Venezuela, and many others. People in these locations use sophisticated tools to circumvent geographic restrictions, including virtual private networks (VPNs), remote desktop services, and IP address rotators. For most of these people, even the modest payments collected from online surveys represent a meaningful amount of money creating the incentives for fraud.

Yea-Saying—A Common Behavior in Online Fraud

One thing revealed by the survey and interviews described above is that fraudulent participants say 'yes' to most survey questions, even when those questions ask about impossible or implausible events. The reason for this behavior lies in how participants qualify for various surveys in the market research ecosystem.

Most market research surveys target people who fit a specific profile. Companies want to hear from consumers who use specific products or services, fall into certain demographic categories, or engage in specific behaviors. To identify these people, researchers use screening questions.

Consider how this works. A soft drink company might want feedback from people who regularly consume their products. Their survey begins with questions like, "Do you drink Coca-Cola?" or "Have you purchased any of the following beverages in the past month?" A little later in the survey, the researchers might ask about specific product usage or preferences. Someone whose

answers do not fit the profile the researchers are looking for is immediately disqualified. The survey ends and the person receives no compensation for their time.

This system creates a powerful incentive to say 'yes.' Participants quickly learn that saying "yes" to screening questions increases their chances of qualifying for the study and earning money. Over time, the behavior becomes habitual regardless of the question's content or the participant's actual experiences. For participants working in click farms this yea-saying behavior is even more extreme because it is explicitly taught to maximize income. Indeed, there is an online cottage industry that teaches people how to commit fraud in online surveys.

The Global Network of Research Fraud

Survey fraud is perpetuated through extensive online communities. Platforms like YouTube, Facebook, Telegram, and Reddit host thousands of tutorials and discussion groups that provide detailed instructions for how to bypass security measures in online panels. Some YouTube channels have tens of thousands of followers and post daily videos explaining how to create convincing false identities, manipulate location data, and pass common screening questions.

One notable channel, called "Survey Help 5.0," is based in Bangladesh. It has amassed 44,000 YouTube followers and nearly 6,000 Facebook followers. This channel posts daily videos showing people how to circumvent security measures in online panels. Some videos provide step-by-step instructions for posing as a U.S. citizen, including detailed guidance on using rented U.S. phone numbers and proxy servers to obtain an IP address in the United States. Another video demonstrates how to create a profile and pose as a Black woman, complete with a fake driver's license and techniques to participate in video surveys specifically recruiting Black female participants.

Social media accounts have built substantial followings by teaching these deceptive tactics. They explicitly instruct followers to maximize survey earnings by claiming to fit multiple target demographics and to respond "yes" to as many screening questions as possible. By indiscriminately answering "yes" to these questions, participants increase their chances of qualifying for studies regardless of whether they meet the criteria. Some people have even taken

to publishing digital courses that teach others how to qualify for lucrative research opportunities like in-depth interviews, focus groups, ethnographic studies, and video research.

Overall, the global network of 5 billion annual surveys contains a heavy dose of fraudulent data from respondents all around the world. This fraud is fueled by economic opportunity and amplified by social media networks that disseminate information for how to circumvent the protection mechanisms that panels have put in place. But fraud may not be limited to one side of the research equation.

In April 2025, the U.S. District Attorney's Office in New Hampshire filed an indictment accusing two market research panels of intentionally directing surveys intended for U.S. participants to people in click farms outside of the U.S (U.S. Attorney's Office, 2025). By fabricating large amounts of survey data, these platforms are alleged to have defrauded their customers out of more than \$10 million. While it is rare for panels to be complicit in the proliferation of fraud, this indictment underscores the need to be vigilant about data quality online.

The Rise and Fall of Mechanical Turk

The section above describes data quality of market research panels, but the story of data quality on Mechanical Turk reveals how an otherwise high-quality platform can have its reputation sullied by fraud almost overnight.

MTurk was created in 2005 as a platform where people could solve problems that computers could not handle efficiently. The platform connected “requesters” (people who need work done) with people MTurk refers to as “workers.” It did not take long for academic researchers to suggest MTurk might be good for research. At the time, most behavioral science studies collected data from undergraduate students. As a result, MTurk emerged at an opportune time, and it provided researchers with access to diverse participants at an affordable rate without requiring technical expertise (see Litman and Robinson, 2020; Moss et al., 2024).

Following the publication of an influential paper, MTurk was rapidly adopted by the scientific community (Buhrmester et al, 2011). Numerous papers showed that data quality on MTurk was very good (e.g., Buhrmester et al., 2011; Litman et al, 2015; see Bohannon, 2016). Indeed, the

data from people on MTurk was often better than that from undergraduate students (Hauser and Schwarz, 2016). As a result, MTurk quickly replaced the undergraduate subject pool as the main source data in the social and behavioral sciences. By 2016, MTurk data represented nearly 50% of studies reported in top psychology journals (Zhou & Fishbach, 2016). And by 2018, MTurk had been cited in more than 1,000 different journals (Buhrmester et al., 2018).

However, starting in 2018, a dramatic shift occurred. Researchers began to see unusual patterns in their data, including inconsistent demographic information, nonsensical responses to open-ended questions, and unprecedented failure rates on attention checks (Bai, 2018; Ryan, 2018). These problems emerged almost overnight and caused widespread concern. The issue was so big that news outlets like *Wired* and the *New Scientist* ran headlines such as “Bots on Amazon’s Mechanical Turk are ruining psychology studies” (*Wired*, 2018) and “A Bot Panick Hits Amazon’s Mechanical Turk” (*New Scientist*, 2018).

When more detailed analyses of the problem were conducted, however, researchers found little evidence of bots (e.g., Moss & Litman, 2018). Instead, a significant portion of MTurk workers appeared to be providing low-quality data from outside of the U.S., in some cases reaching 30-50% of responses (Chmielewski & Kucker, 2019; Dennis et al., 2020).

Research revealed that international participants had circumvented MTurk's geographic restrictions using virtual private networks (VPNs) and other technical tools (Dennis et al., 2020). Conversations on social media revealed that people in India and other countries were buying access to accounts created by people in the U.S. (Figures 10.3 and 10.4). An analysis of questionable participants' IP addresses revealed that many supposedly U.S.-based respondents were from countries like Venezuela, India, and Eastern European nations (Moss & Litman, 2018). These participants provided data with a distinct pattern, often including impossibly fast completion times, inconsistent demographic information across studies, and implausible response combinations (Kennedy et al., 2020).

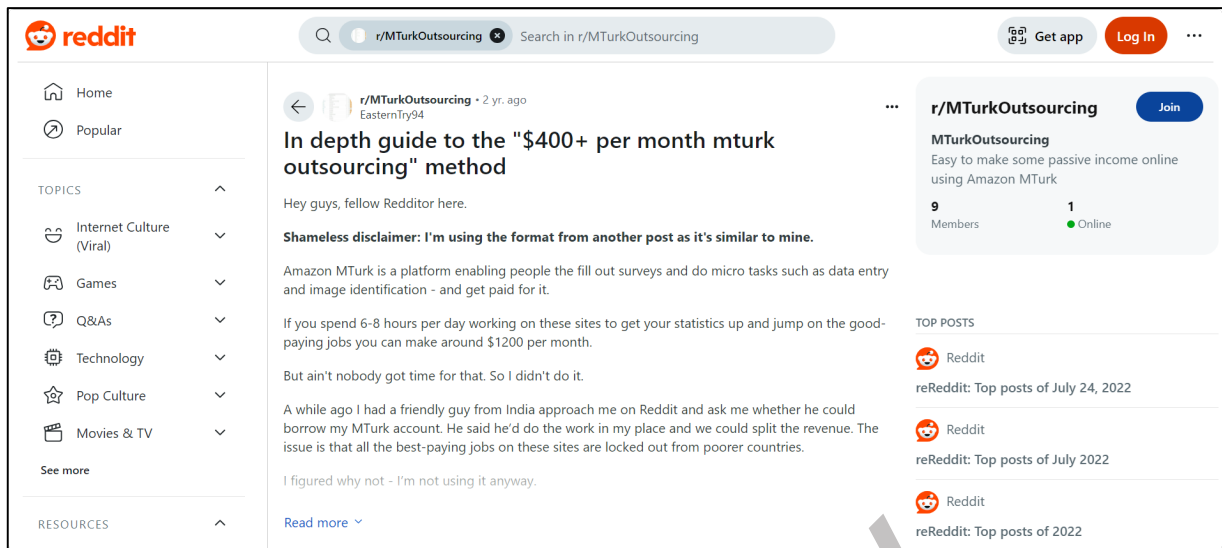


Figure 10.3. A Reddit post explaining how to sell participant credentials to fraudsters outside the U.S.

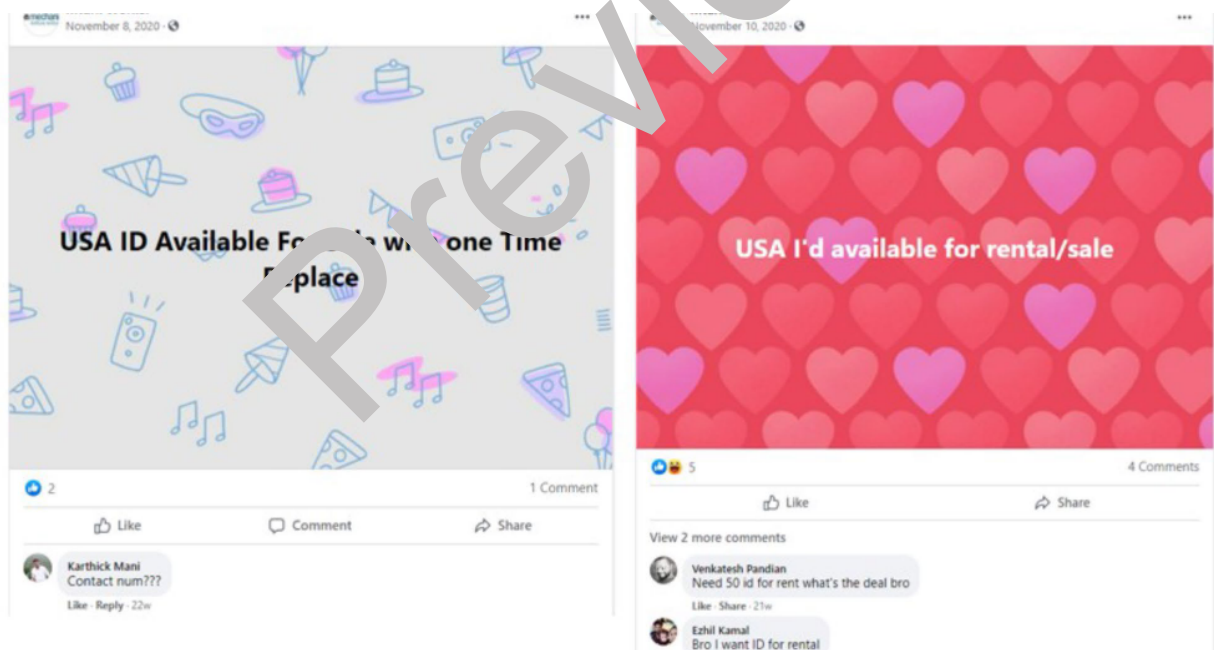


Figure 10.4. A Facebook post seeking to rent a participant ID on MTurk.

Eventually, a large study found that 65,000 out of 165,000, or about 40%, of MTurk workers provided unreliable data (Hauser et al., 2023). While the source of the data quality issue was

often described as “bots,” research painted a different picture. Evidence pointed to human fraud coming from outside of the United States (see Kennedy et al., 2020a, b; Litman et al., 2021). Indeed, it did not take long to see that the data quality issues on Mechanical Turk were an extension of the same problems that had existed for years on market research panels and panel aggregators.

The fall from grace of Mechanical Turk has led to several shifts in online data collection practices. First, specialized platforms emerged to replace Mechanical Turk. These platforms implemented rigorous data quality verification procedures. Sites like CloudResearch Connect and Prolific built their systems with data quality as a central focus, reducing the prevalence of fraudulent data to the low single digits compared to the 30-40% or more found on typical market research panels (Stagnaro et al., 2024).

Researchers also put more emphasis on behavioral and technical solutions that could be implemented within studies regardless of which platform was used. Behavioral solutions including attention checks, instructional manipulation checks, and data validation approaches became standard practice (e.g., Arndt et al., 2022; see Chapter 13). Technical solutions, such as CloudResearch’s Sentry were developed to identify suspicious IP addresses and patterns among participant accounts, helping researchers filter out potentially fraudulent participants before they entered studies (Litman et al., 2021).

The rise and fall of Mechanical Turk as the dominant platform for behavioral research illustrates both the promise and the perils of online data collection. While the platform revolutionized access to research participants and dramatically accelerated the pace of behavioral science, it also demonstrated that data quality cannot be taken for granted. The lessons learned from this period continue to shape how researchers approach online data collection, with increased emphasis on verification, quality screening, and transparent reporting of data quality measures.

High Quality Online Platforms

We began Chapter 9 with a study conducted by Open AI and the MIT media lab in which they followed 1,000 people over 30 days, examining how interactions with ChatGPT affected loneliness. The study was run on CloudResearch Connect, the same platform used throughout this book. Before 2018, the study probably would have been run on Mechanical Turk. But

because of the data quality problems described above, researchers have largely replaced MTurk with Connect, and other platforms like Prolific, that research shows provide substantially higher quality data (e.g., Stagnaro, 2024).

Crowdsourcing platforms like Connect have been built with data quality as a focus. For that reason, they stand out in the online research ecosystem. While MTurk and the typical market research panel may contain 30-40% fraudulent responses, studies conducted on specialized academic platforms like Connect contain less than 5% problematic respondents and often even less. This improvement stems from a vetting process during participant registration, continuous monitoring of people's behavior on the site, and sophisticated detection systems designed to identify response patterns associated with fraud.

For example, Connect requires people to complete an intense onboarding process and submit a valid photo ID to continue taking studies. It also evaluates people's activity after they are on the platform and monitors social media for conversations related to fraud. Finally, Connect solicits feedback from researchers when they detect problematic data. Thanks to these procedures, platforms like Connect and Prolific provide significantly better data quality than MTurk or panel aggregators such as Lucid (Stagnaro et al., 2025).

However, current success does not ensure future outcomes. Even platforms like Connect need to be monitored for data quality to remove potential problematic data. Additionally, as mentioned in Chapter 9, there is often a tradeoff between different participant sources and researchers will sometimes need a market research platform to meet different research objectives. In the next chapter, we examine ways to identify and remove problematic data when using online platforms.

Before that, however, Module 10.2, shows how problematic data show up differently in descriptive, correlational, and experimental studies, how they can harm research findings in each type of study, and we begin to illustrate the various techniques that have been successful in finding and removing problematic respondents in past studies.

How Data Quality Affects Research

Module 10.2

Learn how low-quality data distorts descriptive, correlational, and experimental research

Researchers do not want fraudulent data in their studies, but why exactly? What effect does low quality data have on research findings? In this module, we explore how low-quality data distort descriptive, correlational, and experimental research findings.

How Bad Data Affect Descriptive Research.

In June of 2020—three months into the COVID-19 pandemic in the U.S.—researchers from the Centers for Disease Control and Prevention (CDC) published a study claiming that 40% of Americans were engaged in dangerous cleaning practices. Most concerning, about 10% of people reportedly drank or gargled bleach, 4% drank or gargled tap water, and 4% drank or gargled other household cleaners to prevent Covid-19 infection (Gharapour et al., 2020). You may recall, the start of the pandemic was a scary time. Yet, if these numbers were accurate, then tens of millions of people were not just misusing cleaning products but doing things that defied common sense.

Perhaps unsurprisingly, an attempt to replicate the CDC's results led to a very different picture (see Litman et al., 2023). This replication used the same questions and sampling procedures as the CDC but included data quality checks to identify potential fraud and inattention (the next chapter will describe these checks in more detail).

In the replication study, researchers separated participants into two groups: those who failed the screening measures (“unreliable”) and those who passed (“reliable”). Across two studies, the initial results matched the CDC's findings. However, when the researchers examined the data by group, a different picture emerged.

As Figure 10.5 shows, dangerous cleaning behaviors were reported exclusively by unreliable participants. Reports of drinking bleach or other dangerous chemicals came exclusively from participants who failed the quality checks.

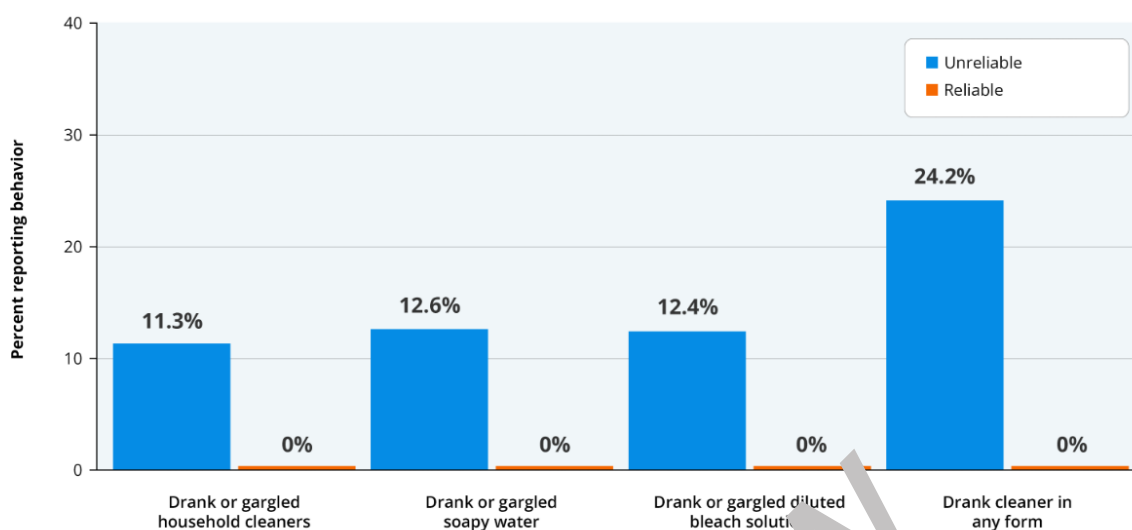


Figure 10.5. After screening for data quality, almost no reliable participants reported dangerous cleaning behaviors. Nearly all such reports came from unreliable participants. From Litman et al. (2023).

What does this reveal about how poor data quality distorts descriptive research? Descriptive studies typically report percentages or averages (point estimates) to characterize populations. When data includes responses from inattentive or fraudulent participants who engage in systematic yea-saying, the estimates become illusory. In effect, when twenty percent or more of respondents are systematically agreeing with most items, evidence can be found for the existence of anything.

Many of the polling results listed earlier were befallen by a similar fate. We know this because follow-up studies, with proper data quality measures, found much smaller numbers or entirely non-existent effects (e.g., Hartman et al., 2023; Holliday et al., 2024; Litman et al., 2023; Mercer et al., 2024). Yet the sensational claims about Americans drinking bleach, Millennials denying the Holocaust, or Black people saying it isn't okay to be White made headlines because they were shocking—and because the researchers failed to account for data quality problems.

How Bad Data Affect Correlational Studies

We have seen how poor data quality distorts descriptive research by inflating point estimates. But how do these same problems affect correlational research?

Remember that a correlation shows how two variables relate to each other. As described in Chapter 5, when one variable increases or decreases, the other does too in a predictable direction. When participants engage in yea-saying—systematically agreeing with questions—they artificially increase the correlation between measurements. This can make unrelated variables appear correlated or make weak relationships look stronger than they really are.

Let's look at an example. Figure 10.6 shows the correlation between people's education and social anxiety from a study by Chandler et al., (2020). Previous research found these variables have a small negative correlation—as education increases, social anxiety tends to decrease. This makes sense given that succeeding in higher education requires navigating complex social environments. However, the correlation in Figure 10.6 is not negative.

In the figure, green dots represent “reliable” participants who passed the quality checks, while red dots represent “unreliable” participants who failed these checks.

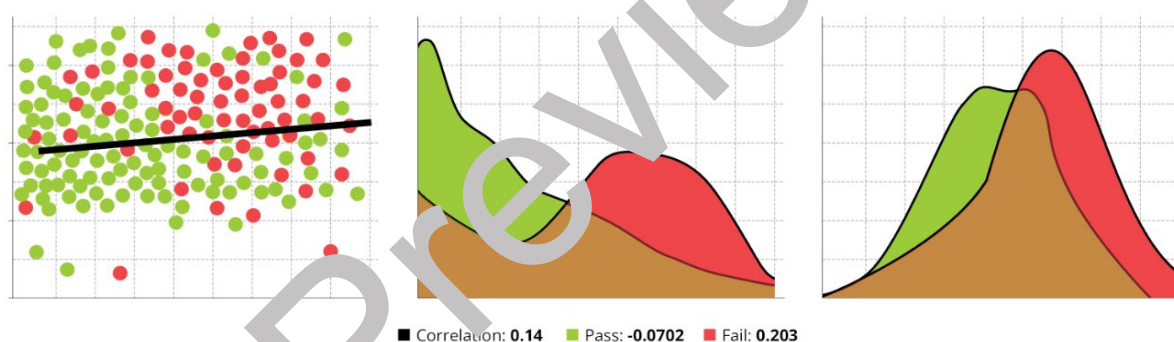


Figure 10.6. Correlation between education and social anxiety showing different patterns among reliable participants (green) and unreliable ones (red). Unreliable participants' tendency to agree with most questions created a false positive correlation, while reliable participants showed the expected slight negative correlation. The red in the histograms reveals how unreliable participants are driving the positive correlation. Data from Chandler et al., 2020.

Looking at the overall correlation in the sample ($r = .14$), we see a positive relationship that contradicts previous findings. But when we separate the data, two different patterns emerge. Among reliable participants, the correlation is slightly negative ($r = -.07$), matching previous research. Among unreliable participants, it is moderately positive ($r = .20$). Why the difference?

The scatterplot on the left shows the clear separation between reliable and unreliable participants. The middle and right images reveal the cause: unreliable participants consistently gave higher ratings on both social anxiety (middle) and education (right). Their yea-saying shifted both distributions to the right, creating an artificial positive correlation between variables that should be negatively related.

Something similar happened within the CDC study we examined earlier as well. After asking about cleaning practices, researchers measured negative health outcomes like skin irritation, dizziness, headaches, and breathing problems. Among reliable participants, there was no correlation between dangerous cleaning behaviors and health problems ($r = .03$). But among unreliable participants, there was a highly significant correlation ($r = .8$). When analyzed together, the unreliable responses inflated the overall correlation, leading CDC researchers to wrongly conclude that dangerous cleaning behaviors were associated with health problems.

False correlations pose a serious risk to research. When researchers find surprising correlations that do not match expectations, those results often face scrutiny. But when inflated correlations match what researchers predict, they rarely question the findings. Instead, they see the correlation as evidence supporting their hypothesis. This means flawed findings that confirm expectations are more likely to be published and enter public discourse. In fact, several examples of this phenomenon exist.

Consider, once again, the Add Health study. After identifying participants who lied about immigrating to the U.S., researchers reanalyzed their data. Table 10.2 tells the story. The first column compares true immigrants to U.S.-born adolescents, showing small differences that generally favored immigrants. The second column compares adolescents who falsely claimed to be immigrants against U.S.-born adolescents, showing large differences suggesting immigrants face more problems.

Participants who lied about being immigrants reported far more negative outcomes—more trouble in school, drinking, emotional distress, health problems, and fighting. When researchers analyzed all the data together, they concluded that immigrant students struggled more than U.S.-born students. The finding seemed logical given the challenges of adjusting to a new country, but it was based on false data.

Outcome Variable	True Non-U.S. Born (n = 863) vs. True U.S. Born (n = 11,550)	False Non-U.S. Born (n = 176) vs. True U.S. Born (n = 11,550)
School grades (+)	-0.03	-0.32
School trouble	0.03	0.30
Positive school feelings (+)	0.01	-0.64
Skipping school	0.02	1.33
Smoking	-0.29	0.63
Drinking	-0.24	1.07
Drunk	-0.26	1.24
Self-esteem (+)	0.01	-0.49
Emotional distress	-0.10	0.60
Future hope (+)	-0.03	-1.32
Health problems	-0.36	0.67
Physical problems	-0.15	1.62
Sickness	-0.04	0.90
Fight	0.21	1.15
Lie to parents	0.21	0.59
Mean of absolute effect sizes	0.13	0.80

Table 10.2. Comparison of true U.S.-born adolescents, true immigrants, and those who falsely claimed immigrant status. False immigrants reported dramatically worse outcomes, artificially inflating apparent differences between immigrant and U.S.-born youth. From Fan et al. (2006).

Research on LGBTQ youth has shown a similar problem. Since relatively few adolescents identified as LGBTQ in the past, even a small number of mischievous responders could significantly distort findings. Some studies confirm this problem (Robinson & Esplage, 2011, 2012, 2013; Cimpian & Timer, 2020; Savin-Williams & Joyner, 2014). While LGBTQ youth do face higher risks for some issues like suicidal thoughts and bullying, these risks appear much larger when mischievous participants remain in the data. Other supposed risk factors like drug abuse and fighting disappear almost entirely when mischievous participants are removed.

Finally, another example appears in Figure 10.7, showing the correlation between Facebook use and depression. Once again, we see clear separation between reliable participants (green dots) and unreliable ones (red dots). The correlation among reliable participants is small ($r = .11$), while among unreliable participants it is moderate ($r = .35$). Analyzing all data together yields an inflated correlation ($r = .32$).

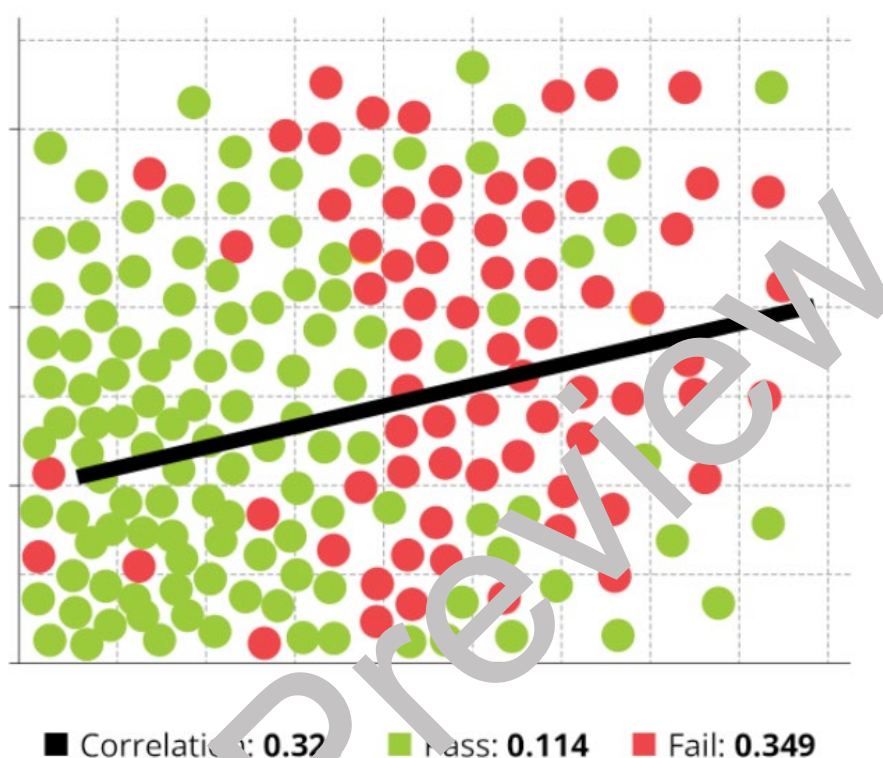


Figure 10.7. Facebook use and depression correlation, showing how unreliable participants (red) significantly inflated the relationship compared to reliable participants (green). Data from Chandler et al., 2020.

A researcher finding this inflated correlation would likely publish it without question, contributing to concerns about social media's impact on mental health. The problem is that this finding largely results from low-quality data. And this is how bad data threatens correlational studies—by giving researchers a false understanding of how variables relate to each other. These artificial relationships can mislead both scientists and the public, potentially driving misguided interventions and policies.

How Bad Data Affect Experimental Studies

We have seen how poor data quality distorts descriptive studies and creates false correlations. In experimental research, these same problems lead to a different issue: biased estimates of effect size.

An effect size tells researchers how big the difference is between experimental conditions (see Chapter 7 if you need a refresher on experimental design). When a manipulation strongly affects people's thoughts, feelings, or behaviors, the effect size is large. When the difference between conditions is small, so is the effect size.

The famous Trolley Dilemma provides an example of how bad data quality can affect experimental results. In the Trolley Dilemma, participants imagine a train barreling down the tracks toward five people who are tied up and unable to move. In one version, participants can pull a lever to divert the train to a different track where it will hit and kill just one person. The critical question is whether participants would pull the lever to save five people or let the train continue its path.

If you are like most people, you feel there is something wrong with letting five people die over just one. So, as unpleasant as it may be, 70–80% of people across countries and cultures choose to pull the lever and save the lives of five people at the expense of one (Awad et al., 2020). But as in all experiments, there is another condition.

In the second version of the trolley dilemma, people can push a large man off a bridge above the tracks to stop the train, saving the five trapped people but killing the man who was pushed. As in the original version, the critical question is: what would participants do?

If you are like most people, you are less certain about this scenario. In fact, the number of people willing to save five people at the expense of one often falls by more than half across these scenarios (Awad et al., 2020). The difference between these conditions is so reliable that researchers consider it a human universal. When people are paying attention and responding honestly, this effect appears again and again.

The reliability of the Trolley Dilemma makes it useful for examining data quality. In one study, researchers compared two groups: a "reliable" group of participants who had passed a prior screening process, and a "unreliable" group who had failed this screening.

Figure 10.8 shows the results. Reliable participants showed the expected pattern—they were much more willing to pull the lever than push the man. But unreliable participants showed no difference between conditions. About 90% in both conditions said they would act to stop the train—a pattern that contradicts decades of research. These unreliable participants simply agreed with whatever option was presented (i.e., said Yes, I will pull the lever; Yes, will push the man), regardless of the moral implications.

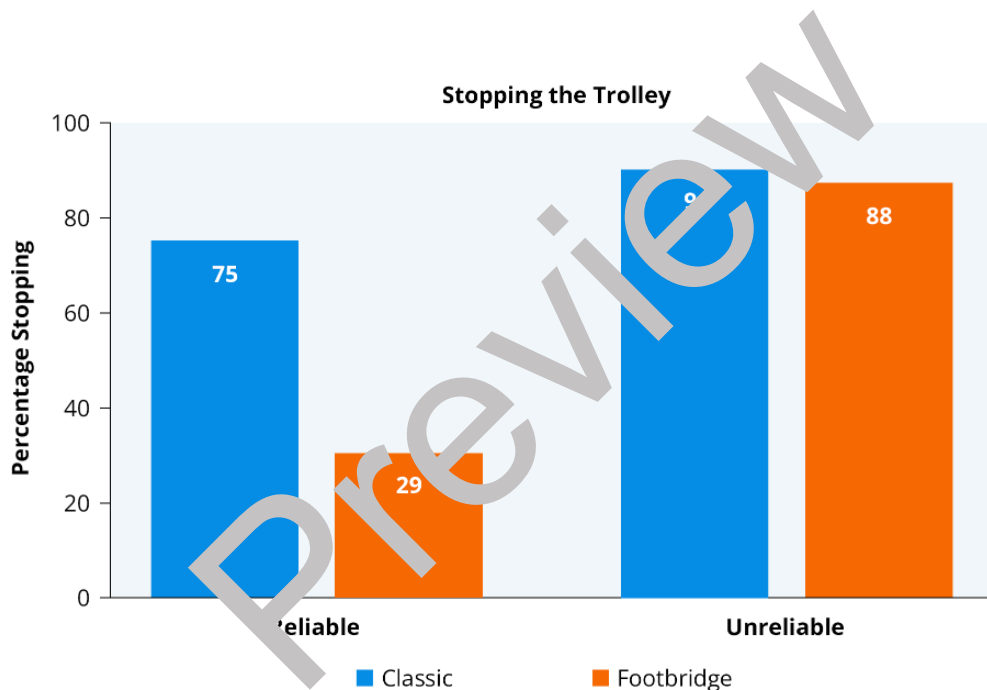


Figure 10.8. Trolley dilemma responses by participant group. Reliable participants showed the expected pattern (more willing to pull lever than push man), while unreliable participants showed uniformly high willingness in both conditions, contradicting established findings.

Another study used something called the "soda task" to demonstrate how data quality affects experimental results (Hauser et al., 2023). In this task, participants indicate how much they would pay for a soda on a hot day at the beach. Some participants are told the soda comes from a run-down grocery store, while others are told it comes from a fancy resort.

Researchers tested this scenario with three groups: participants with good data quality (who passed prior screening), participants with bad data quality (who failed screening), and a mixed group who had not been screened.

As Figure 10.9 shows, among participants with good data quality, there was a large effect—people were willing to pay significantly more for the resort soda than the grocery store soda. In the mixed quality group, the effect was present but much smaller. Finally, in the bad quality group, the effect disappeared completely.

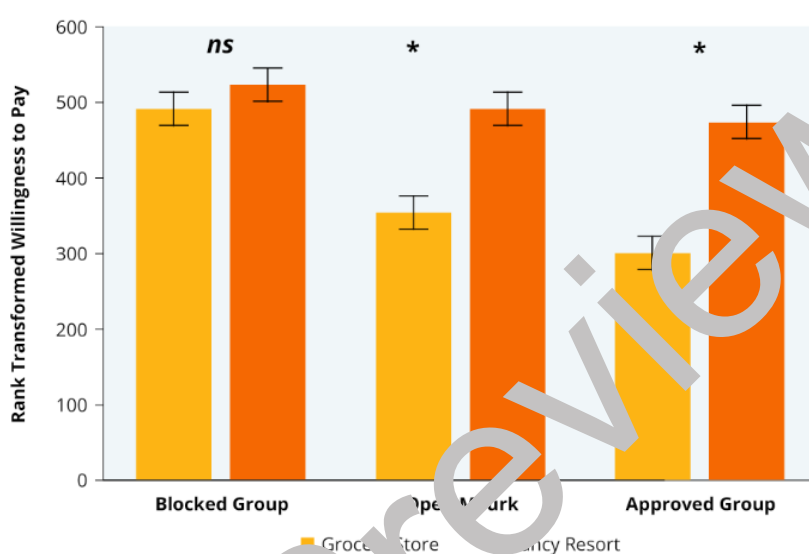


Figure 10.9. Soda task results across three data quality conditions. The price difference between resort and grocery store sodas was large with good quality data, reduced with mixed quality data, and disappeared entirely with poor quality data.

This pattern reveals the main consequence of data quality problems in experiments. If the responses are systematic, such as consistent yea-saying in the Trolley Dilemma, they will bias the effect in a specific direction. If the responses are random, they create noise that can overwhelm the effect researchers are looking for, even when that effect truly exists. Poor quality responses essentially dilute or distort the impact of experimental manipulations.

This has real consequences beyond the lab. If researchers conclude an intervention has a smaller effect than it truly does, promising treatments or policies might be abandoned. Conversely, if

researchers cannot detect effects that truly exist, they might wrongly conclude that their hypothesis is incorrect.

Across all three research types—descriptive, correlational, and experimental—data quality problems lead to fundamentally misleading conclusions. The good news is that researchers can implement effective strategies to detect and address these issues, as we will explore in the next chapter.

Summary

Data quality is fundamental to the integrity of behavioral research, as this chapter has demonstrated through concrete examples. When participants provide low-quality data—whether through inattention, yea-saying, or deliberate misrepresentation—consequences affect all research methods.

In descriptive research, poor data quality distorts point estimates, creating false impressions about the prevalence of attitudes and behaviors, as seen in the CDC bleach drinking study. In correlational research, participant behaviors like yea-saying artificially inflate relationships between variables, as demonstrated in the Facebook depression and education-anxiety examples. In experimental studies, low-quality responses dilute effect sizes and can obscure real differences between conditions, as shown in the Trolley Dilemma and soda pricing experiments.

While random noise creates challenges, systematic bias poses the greater threat because it consistently distorts findings in predictable ways. This systematic distortion can create relationships where none exist, inflate weak associations into strong ones, or mask experimental effects entirely. Most concerning is when these distortions align with researchers' expectations, making them less likely to be scrutinized.

The Add Health study and other examples throughout this chapter illustrate that data quality problems can affect even the most methodologically rigorous research. The good news is that researchers can implement effective countermeasures. In the next chapter, we present practical strategies to identify and address data quality issues, enabling more accurate and reliable behavioral research.

Research Activity 10.1: Data Quality in the Wild

This chapter has shown how problems with data quality can ruin research. In the guided project below, you will get a deeper look at how low-quality data can mislead not only researchers but the media and the public. You will also see first-hand how data quality problems are detected, which is a preview of the next chapter on data quality solutions. After completing these assignments, you will have a better understanding of:

1. The nature and prevalence of data quality problems in online research and how they manifest themselves in surveys.
2. How researchers and research consumers can be misled by low quality data when studies do not have proper data quality practices in place.
3. The tools and measures that behavioral scientists use to detect data quality problems and how to verify important information that participants share online.

Background

To clarify how problems with data quality affect research, you will dig into what amounts to a case study. First, we will ask you to read a short study published by researchers at the Centers for Disease Control and Prevention (CDC), claiming to show that 39% of Americans were engaged in dangerous cleaning practices during the COVID-19 pandemic, including about 4% who said they drank bleach. After reading the original report, you will watch a short video showing how the study became international news covered by hundreds of media outlets. Finally, you will read the research article that showed the CDC's numbers were inflated by fraud.

The goal of this assignment is for you to see how, even when researchers, journalists, and others try their best, poor data quality can create a misleading understanding of people's thoughts, feelings, and behaviors.

Assignments

1. Read the report released by the Center's for Disease Control and Prevention in June 2020.

The report, titled “Knowledge and Practices Regarding Safe Household Cleaning and Disinfection for COVID-19 Prevention — United States, May 2020” (Gharpure et al., 2020), was released a few months into the COVID-19 pandemic and sought to describe dangerous cleaning behaviors people were engaged in to protect against the virus. After reading the brief article, write down your initial thoughts in a document.

What kind of sampling methods did the authors use, and what kind of research were they conducting? Did the article mention anything about data quality? What were the rates of different dangerous behaviors that people reported? How well done is the report in your mind? Are there reasons to be concerned?

2. Watch the video titled “News coverage of Americans drinking bleach.”

After reading the article, watch the video detailing the media attention the CDC findings received. You can find the video on the Research in the Action YouTube page under the title “Did People Really Drink Bleach? – Media Coverage.” Within a few weeks of being published, the CDC paper had been reported on by hundreds of media outlets worldwide. It was also discussed in thousands of social media posts, the topics of blogs, and apparently even cited in two Wikipedia articles.

As you watch the video, ask yourself how well the news articles covered the research and why. Evaluate the coverage the article received. Did the news outlets get the story “right”? Were the findings simplified too much or sensationalized?

3. Read the article, “Did people really drink bleach to prevent COVID-19?”

The research article written by Litman et al., (2023) describes how behavioral scientists discovered the CDC results were inflated by survey fraud. After you read the article, summarize your reactions to it. Describe how the researchers established that the CDC results were due to fraud, how they detected data quality issues in their study, and what the results looked like once the data quality issues were corrected. Describe what measures the researchers used to find low quality data. Do you find their account convincing?

Additional Readings

- Berry, C., Kees, J., & Burton, S. (2022). Drivers of data quality in advertising research: Differences across MTurk and professional panel samples. *Journal of Advertising*, 51(4), 515-529.
- Chandler, J., Paolacci, G., & Hauser, D. J. (2020). Data quality issues on Mechanical Turk. In L. Litman and J. Robinson (Eds.) *Conducting Online Research on Amazon Mechanical Turk and Beyond* (95-120). Sage Academic Publishing. Thousand Oaks: CA
- Jaffe, S. N., Moss, A. J., Rosenzweig, C., Gautam, R., Robinson, J., & Litman, L. (2025). The bots ruining social science aren't bots at all. Manuscript under review.
https://osf.io/wr8ds_v1
- Moss, A., & Litman, L. (2018). After the bot scare: Understanding what's been happening with data collection on MTurk and how to stop it. Retrieved February, 4, 2019.
<https://www.cloudresearch.com/resources/blog/after-the-bot-scare-understanding-whats-been-happening-with-data-collection-on-mturk-and-how-to-stop-it/>
- Stagnaro, M. N., Druckman, J., Berinsky, A. J., Prechar, A. A., Willer, R., & Rand, D. G. (2024, February 22). Representativeness versus Response Quality: Assessing Nine Opt-In Online Survey Samples. <https://doi.org/10.31234/osf.io/h9j2d>

Chapter 11: Data Quality Solutions

If you conduct enough research, you will encounter problems with data quality. Sometimes these problems are caused by people who do not pay attention, speed through the survey, or misrepresent themselves. Often, the problem is fraud. As discussed in Chapter 10, the percentage of questionable responses across online sources often ranges from 20-40% (e.g., Weber, 2023). Thus, while the degree of the problem may vary from study to study and from one sample source to another, the problem itself is endemic to online research.

In this chapter, you will learn about effective techniques for identifying and removing problematic respondents from a study. As you learned in the last chapter, low quality data can produce not only misleading conclusions, but a fundamentally incorrect picture of reality.

In Module 11.1, we will examine how to identify problematic participants and how to validate attention checks with benchmark questions. Among the questions we will explore are: how do researchers measure data quality? How do they decide which participants provided honest answers? How can researchers be confident their measures of quality are accurate and what kinds of tradeoffs must they consider when determining the appropriate threshold for removing respondents? To explore these questions, we will walk through an extended example from a large-scale study that examined data quality online (Reavey et al., 2024).

In Module 11.2, we will review the various types of attention checks commonly used in behavioral research. We will examine different approaches to identifying low quality respondents, like attention checks, comprehension checks, yea-saying questions, and open-ended validation, comparing their strengths and limitations. We will examine best practices for writing your own attention check questions and introduce a website, called SurveyDefense, that can be used to generate different kinds of attention checks.

Throughout the chapter, we will emphasize a "fit-for-purpose" approach to data quality. Different research questions require different levels of quality control, and you will learn how to make informed decisions about appropriate thresholds for your specific research goals. By the end of this chapter, you will have practical tools to enhance data quality in your online studies.

Chapter 11 Outline: Data Quality Solutions

Module 11.1: Detecting Fraudulent Responses



Reading

Learn how to spot fraudulent participants

- Explore questions that detect yea-saying, a common behavior among fraudulent participants
- Examine evidence for the effectiveness of yea-saying questions
- Learn about tradeoffs when setting thresholds for excluding participants

Module 11.2: Types of Attention Checks



Reading

Explore the questions researchers use to measure attention

- Learn about various types of attention check questions
- Compare the strengths and weaknesses of different items

Module 11.3: Implementing Data Quality Measures in Your Research



Reading

Explore a strategy to protect your online data collection

- Walk through the steps to develop a complete data quality plan for your research

Detecting Fraudulent Responses

Module 11.1

Learn how to spot fraudulent participants

Imagine you have just collected data from 2,000 people online. Before you can analyze the data, you must identify which participants gave thoughtful and honest answers and which ones did not. How would you do it?

The most common approach is to incorporate special questions in the study that detect inattention and fraudulent responding. These questions are often referred to as **attention checks** (Oppenheimer et al. 2009).

There are many different types of attention checks. In Module 11.2 we will review a variety of techniques that researchers use to measure attention. But first, we will focus on the specific case of how researchers detect yea-saying and how they know these questions are effective. The examples we use are from a real study that examined data quality in online research (see Reavey et al., 2024).

Detecting Yea-Saying

In the last chapter, we described the global network of survey fraud and the behaviors people use to siphon money out of online surveys. One of the most common strategies employed by fraudsters is to say 'yes' to nearly every question, what is known as yea-saying. Yea-saying questions take advantage of problematic participant's tendency to say 'yes' by asking about impossible or highly unlikely scenarios. People who agree with several of these items are flagged as unreliable and examined more closely for possible removal from the study.

In the Reavey et al., (2024) study, five questions were included to detect yea-saying (see Table 11.1). The first question asked: "Have you seen any live music shows in Shea Stadium, NY, in the last two years?" Shea Stadium was demolished in 2009, making it impossible for anyone to have attended a concert there recently. Participants who answered 'yes' to this question were either lying or not paying attention.

Question Type	Question Text	Why It Works
Nonexistent venue	"Have you seen any live music shows in Shea Stadium, NY, in the last two years?"	Shea Stadium was demolished in 2009, making "yes" responses impossible
Fictional cruise lines	"Please indicate if you have cruised with any of the following companies within the last two years: a. Sail Thru Vacations, b. Yacht-ify Cruises, c. Pacific Travails LTD, d. Outstanding Ocean Outings, e. None of these."	None of these cruise companies exist, so only "None of these" is correct
Nonexistent products	"Please select which of the following haircare products you have used in the last six months: a. Bunfocore, b. Pleistorene Grow, c. Truefolica Treatment, d. None of the above."	None of these brands exist, making "None of the above" the only correct answer
Fictional restaurants	"Have you recently eaten food from any of these small food chain stores? a. Cheese and Wine Junction, b. Zesty's Tomato Pies, c. Paulina's Routine King, d. Toasty Pita BBQ, e. None of the above."	None of these food chains exist, so "None of the above" is the only correct answer
Highly improbable event	"Have you filed a homeowners insurance claim due to damage from lightning within the past three months?"	Extremely rare event (occurs in <0.0001% of U.S. population)

Table 11.1. Five yea-saying questions from Keavey et al., (2024).

Another question asked participants which of several cruise lines they had traveled with in the past two years. The answer options included names like "Sail Thru Vacations," and "Pacific Travails LTD." None of these businesses exist, meaning there was only one correct answer: "None of these."

Another question in the study asked about something that could really happen but is extremely rare: "Have you filed a homeowners insurance claim due to damage from lightning within the past three months?" Less than 0.0001% of U.S. households experience lightning damage in any three-month period. Other questions asked about non-existent hair care products and restaurant chains. So, how did participants do?

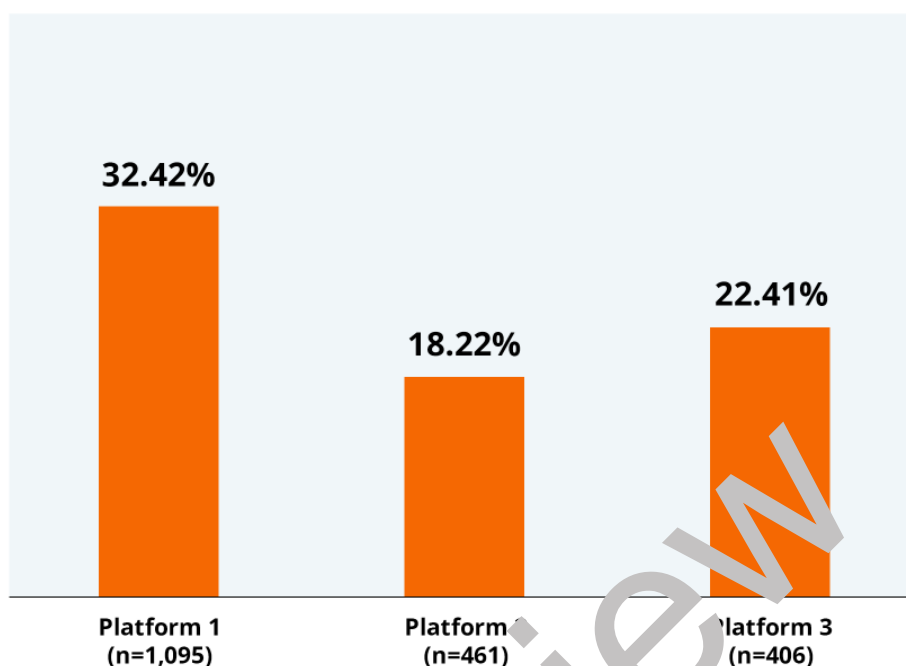


Figure 11.1. Proportion of problematic and fraudulent respondents across three online platforms.

Figure 11.1 shows that between 18% and 32% of participants on each of the three platforms tested said ‘yes’ on at least two out of the five questions. In other words, many participants provide incorrect or unreliable information. But this figure raises several interesting questions.

First, how can we know that the participants who failed these items are bad respondents who should be removed from the study? Perhaps, these are good participants who misunderstood the question or inadvertently provided the wrong answer—what is referred to as a false positive. Second, how can we know whether the yea-saying questions caught all the bad respondents? Perhaps the study included other inattentive or unreliable participants that these questions did not catch—what researchers would call a miss or a false negative. And third, what is the rationale for choosing two out of five as the threshold for identifying problematic responses? Why not use a more lenient (say 3 out of 5 questions wrong) or more stringent threshold (maybe 1 out of 5 wrong) for identifying bad participants? Let’s examine each question one at a time.

Using Benchmarks to Validate Attention Checks

How can we know that the yea-saying questions identified unreliable participants? This is a critical question. If researchers are going to remove 18-32% of their sample, they should be confident they are removing the right people.

The answer comes in what are called **benchmark validation questions**. These are questions where researchers know the percentage of people in the population who should say ‘yes’ to some behavior. By comparing the percentage of ‘yes’ responses to a known population parameter—what is called a base rate—it is possible to tell whether people are responding accurately or not (see Figure 11.2).

In the Reavey et al. study (2024), there were three benchmark questions:

1. "Do you own a Tesla?" (approximately 1% of the U.S. population owns a Tesla).
2. "Have you gone scuba diving within the last 12 months?" (less than 1% do this annually).
3. "Do you follow a vegetarian diet?" (approximately 5% of U.S. adults eat vegetarian).

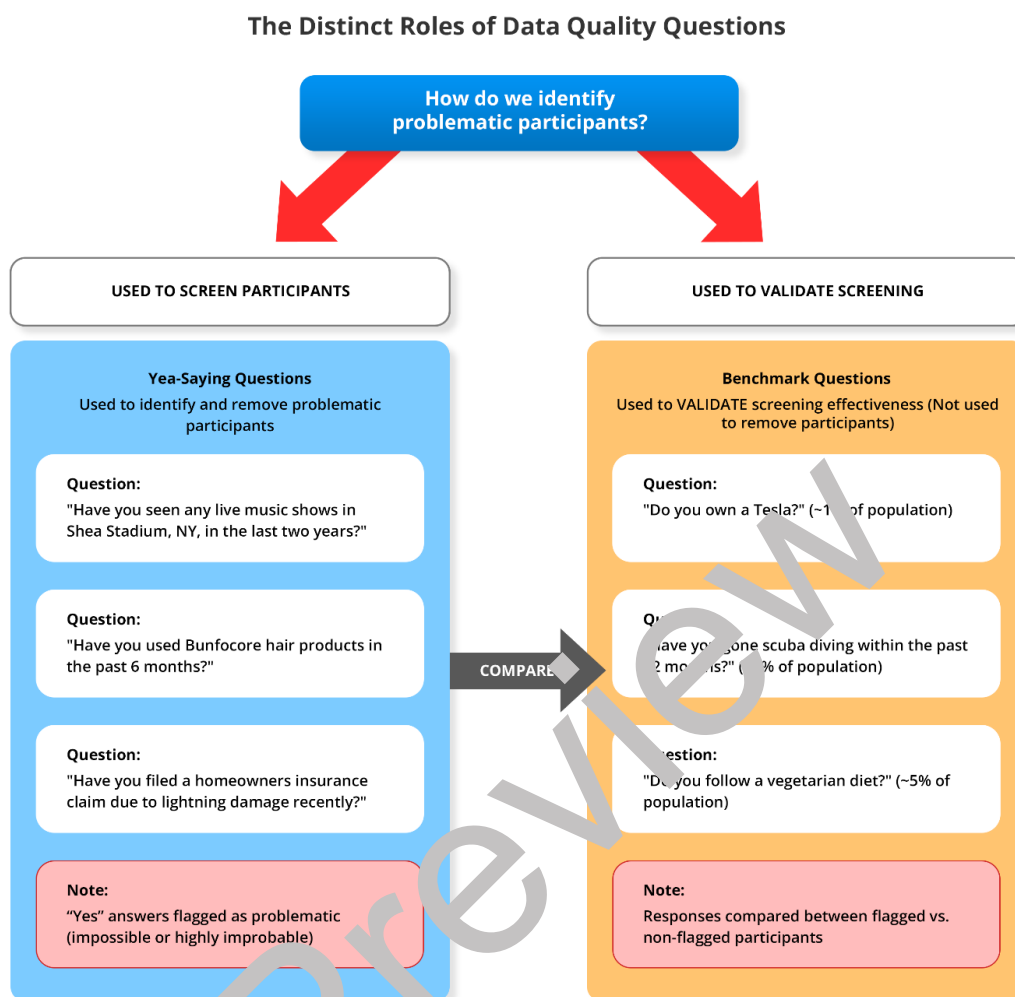


Figure 11.2. Two sets of questions. The first set contains attention checks. The second set contains benchmarking questions. The benchmarking questions are used to validate the attention check questions.

Using these items to validate the yea-saying questions works through a straightforward logic. If the attention checks identify unreliable participants, then the people who fail those items should show a different pattern of response on the benchmarking questions than people who pass the items. Specifically, people who engage in yea-saying should report significantly higher rates of rare behaviors like owning a Tesla, going scuba diving, and eating a vegetarian diet.

As shown in Figure 11.3, this is exactly what happened. Among participants flagged as unreliable, between 40-60% claimed to go scuba diving in the last 12 months—more than 40 times the actual population rate! A similar percentage of participants also reported owning a Tesla (base rate = 1%) and following a vegetarian diet (base rate = ~5%).

In contrast, participants who passed the yea-saying questions (i.e., the "reliable" group) reported values closer to the population base rates: about 4% reported owning a Tesla, between 3 and 5% reported scuba diving, and between 5 and 9% reported eating a vegetarian diet. While not a perfect match to the population rates, these numbers are much closer to reality.

Another way to look at the data—and the way that provides the most compelling evidence for the effectiveness of the benchmarking questions—is to examine how many participants reported *any* of the three rare behaviors. As shown in Figure 11.4, nearly 80% of respondents who failed the yea-saying questions reported at least one of the benchmarking activities. The expected base rate was approximately 7%. Meanwhile, among participants who passed the yea-saying questions, between 9 and 12% reported at least one rare behavior.

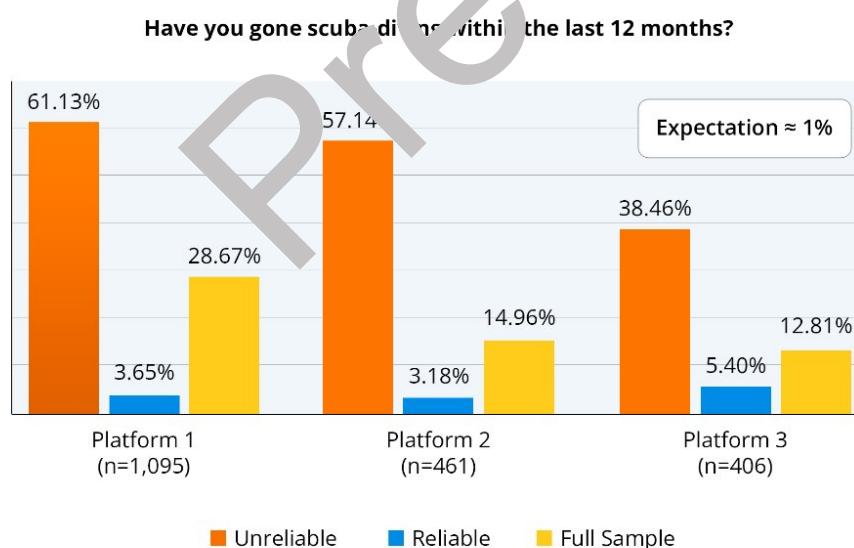


Figure 11.3. Reports of scuba diving among three groups: 1) participants flagged by attention checks, 2) those not flagged by attention checks, and 3) the full sample. The prevalence of scuba diving in the general population is approximately 1%.

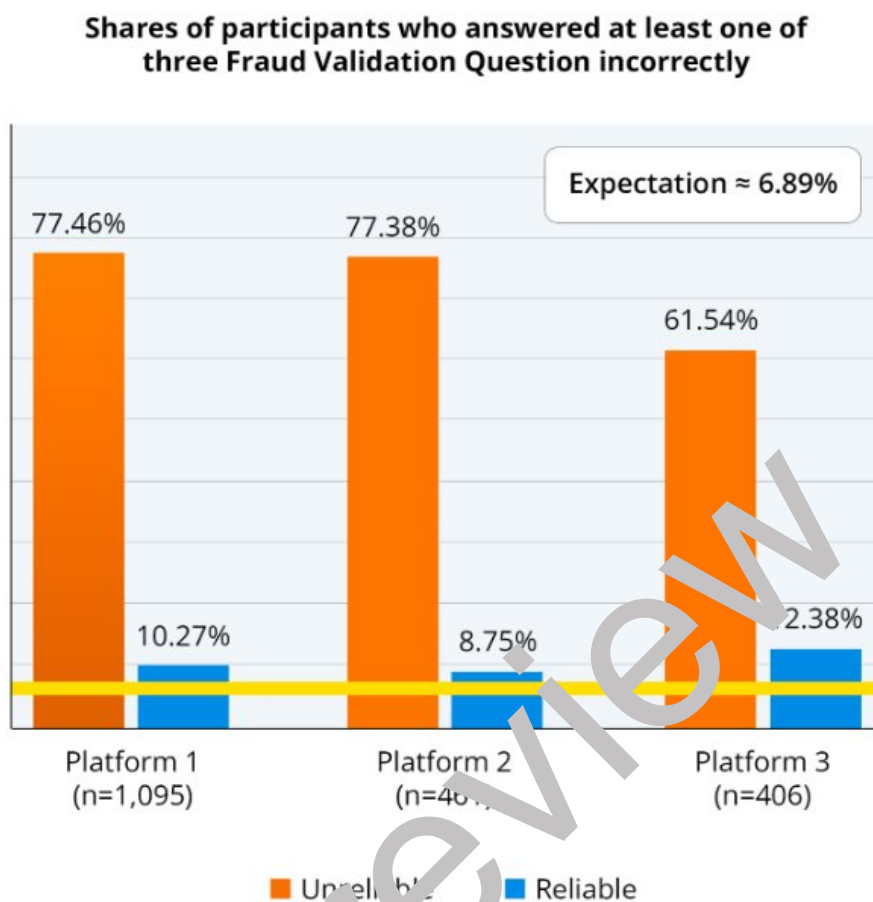


Figure 11.4. Reported rates of engaging in at least 1 of the 3 benchmarking activities (scuba-diving, Tesla ownership, vegetarianism) among three groups: 1) participants flagged by attention checks, 2) those not flagged by attention checks, and 3) the full sample. The prevalence of at least one of the three activities in the general population is approximately 7%.

The dramatic difference between the two groups validates the yea-saying measures and reveals that the benchmark questions help establish two things. First, they establish that people marked as unreliable are, indeed, providing unreliable information not just on one or two yea-saying questions but across the board in the survey. Second, the items establish that people marked as reliable are providing reliable information since their responses to the benchmarking questions are aligned with what is expected in the general population.

At the same time, the benchmarking questions raise other interesting considerations. If participants who pass the yea-saying items are generally honest, then why do they report slightly higher rates of rare behaviors than the population averages? For example, why do 3 to 4% of people report owning a Tesla when the base rate is closer to 1%? It is, of course, possible that some degree of unreliable responding remains in the sample even after removing participants who failed multiple yea-saying items. If so, this observation would lead directly to the next question we will examine: does setting a more stringent exclusion criteria further improve data quality?

Choosing an Exclusion Threshold

Whenever researchers use yea-saying items or attention checks, they must set a threshold for identifying people who should be removed from the analyses. To date, we have examined a threshold of failing at least two out of five yea-saying questions. This approach proved effective at identifying unreliable participants. However, the benchmarking data also showed that participants in the reliable group were not perfectly aligned with the population parameters. This raises the question of whether it would help to use more stringent exclusion criteria?

To explore this issue, consider what happens when we change the threshold from failing two or more questions out of five to failing one or more questions.

As shown in Figure 1.5, the more stringent threshold reduces the percentage of participants reporting rare activities, closer to the population average. Among reliable participants—those who passed all five yea-saying questions—the reported rate of scuba diving drops even further. For example, on platform 2, it drops from 3.2% to 2.1%. As a reminder, the expected population value is approximately 1%. Thus, this suggests a more stringent threshold removed more unreliable respondents who would have been missed with the more lenient threshold.

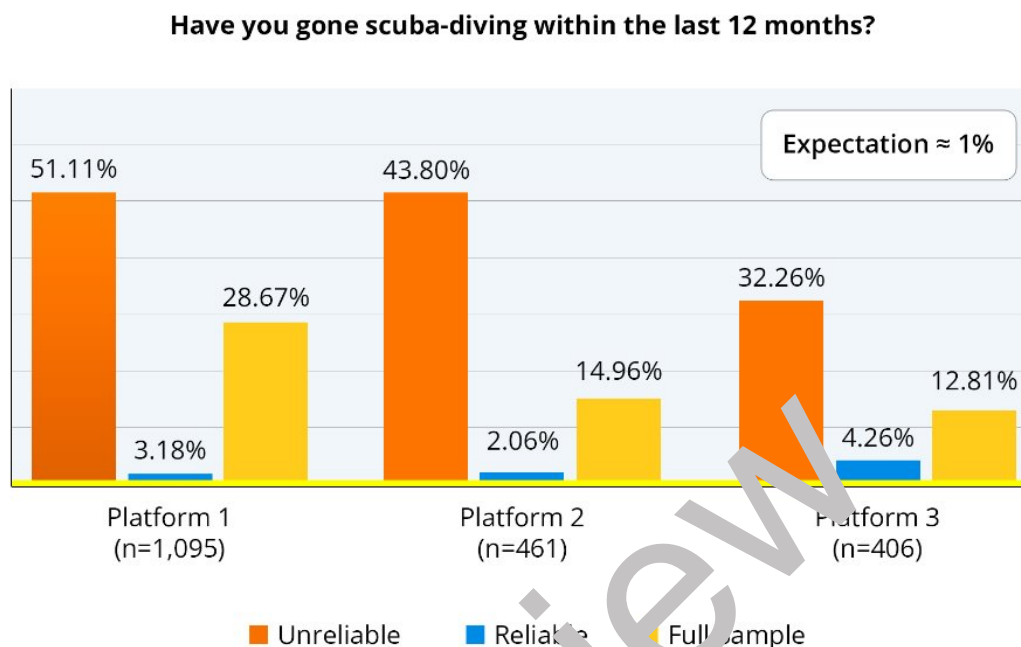


Figure 11.5 Rates of engaging in at least 1 of the 3 benchmarking activities (scuba-diving, Tesla ownership, or vegetarianism) among three groups when a more stringent threshold is applied. Endorsement of even 1 out of 5 yea-saying questions qualifies for being categorized as unreliable or fraudulent.

However, something interesting occurs among people marked as unreliable. With the stricter exclusion criteria, the percentage of unreliable participants who claim to have gone scuba diving also *decreases*—by about 13 points from the earlier analysis. Specifically, 57.1% of respondents reported having gone scuba diving among those who failed at last two yea-saying questions. However, only 43.8% of participants reported having gone scuba diving among those who failed at least one yea-saying question. In other words, it appears that setting a more stringent threshold results in 13% of respondents who were marked as problematic by the yea-saying questions performing perfectly well on the benchmarking questions. The question is, why?

The answer lies in how the two groups are defined. When the more stringent threshold is applied, the unreliable group includes every participant who failed at least one attention check. Many of

these people may be otherwise reliable respondents who made a single mistake, misunderstood the question, or were inattentive for a small part of the survey. They are not, however, *systematically* giving false information, as revealed by the benchmarking questions.

Figure 11.6 illustrates the difference between unreliable and reliable respondents across all three benchmarking questions. You can compare it to Figure 11.4, which uses the more lenient threshold. When comparing the two figures, you will see that the percentage of reliable respondents who claim to engage in the benchmarking activities moves closer to the true population value, 6.9%. At the same time, however, the difference between the unreliable and reliable groups shrinks because the unreliable group contains participants who may have made just a single error in the study.

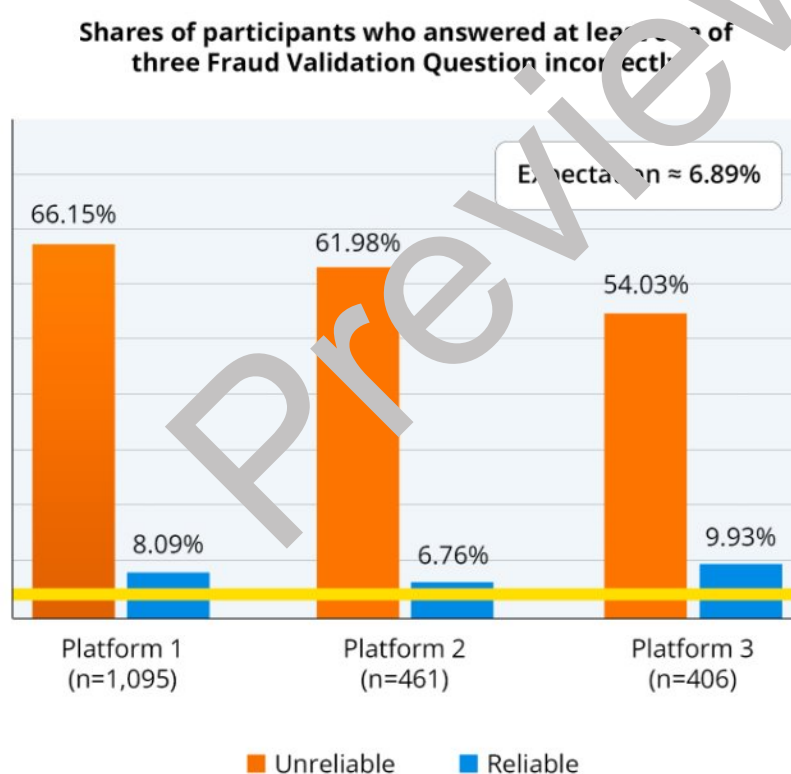


Figure 11.6 Rates of engaging in at least 1 of the 3 benchmarking activities (scuba-diving, Tesla ownership, or vegetarianism) among three groups when using a more stringent threshold. Endorsement of even 1 out of 5 yes-saying questions qualifies for being categorized as unreliable or fraudulent.

The smaller difference between groups demonstrates an important tradeoff in setting exclusion criteria. While being stringent removes more potentially unreliable participants it also increases the number of false positives—reliable participants who are categorized as unreliable. So how does a researcher know where to set the threshold?

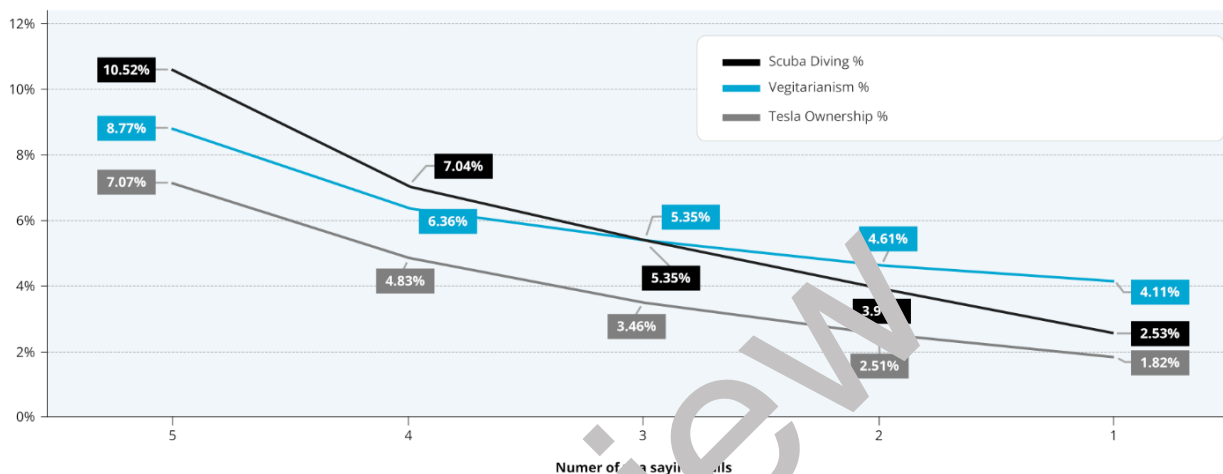


Figure 11.7. How the reported rate of rare behaviors changes as people fail more yea-saying questions.

The answer depends on the study's research objectives. Figure 11.7 provides a visualization of how different thresholds affect data quality. Moving from left to right on the x-axis, the figure shows increasingly stringent thresholds, from only removing participants who failed all five attention checks to removing anyone who failed even a single attention check. The y-axis shows the percentage of participants reporting each benchmark behavior in the cleaned dataset.

What is particularly revealing is how the pattern changes across exclusion thresholds. At the most lenient threshold, the reported rates of benchmarking behaviors are not only inflated but the opposite of what exists in the population. People report scuba diving at a higher rate than following a vegetarian diet, despite scuba diving being much rarer in the population.

Yet as the threshold becomes more stringent, the data corrects itself. When participants who failed two or more attention checks are removed, the pattern starts to align with the population. Following a vegetarian diet is now reported more frequently than scuba diving. But it's only at

the most stringent threshold that the data approaches population benchmarks, with scuba diving reported at 2.51%, Tesla ownership at 1.82%, and vegetarianism at 4.11%—all much closer to their population base rates.

Figure 11.7 illustrates the trade-offs researchers face. Using more stringent exclusion criteria brings the data closer to ground truth but at the cost of removing valid participants. This tradeoff holds several important lessons about data quality.

First, participants are not binary, either all bad or all good. Participants and their data exist on a spectrum of quality. Some percentage of otherwise reliable respondents will lose focus and fail an attention check while otherwise performing well across the rest of the study. Truly unreliable respondents, on the other hand, are likely to fail multiple attention checks.

Second, we advocate for a fit-for-purpose approach to data quality. Different research questions require different levels of data scrutiny.

When the aim of the research is to examine the association between variables or test an experimental effect, a few percentage points of unattentive data are unlikely to threaten validity. For these studies, you can often use a threshold like failing two out of five attention check questions. Indeed, Reavey et al., (2024) found their expected effects using this threshold. The goal of this more lenient threshold is to remove the participants who are almost certainly unreliable while retaining those who may have made an occasional error.

If, however, the aim of the research is to describe people's behavior or to uncover the prevalence of rare events—like the CDC study on dangerous cleaning practices—a more stringent exclusion threshold will be required. This is because even a small percentage of unreliable responses can dramatically inflate the prevalence of rare behaviors. So, to protect your data you might use a threshold of failing one out of five questions, even though it carries a higher risk of false positives. Yet no matter how strict your threshold is some unreliable participants may slip through. Therefore, when the goal of your research is to estimate the prevalence of rare behaviors, we recommend another step that goes beyond asking attention check questions: verifying people's responses with open ended questions.

Open-Ended Validation

Whenever you conduct a descriptive study that aims for a high level of precision in its point estimates, we recommend following up on people's responses to multiple choice questions with an open-ended item that asks them to describe their experience. The value of this approach can be seen in the study described in Chapter 10, relating to the claim that people drank bleach during the COVID-19 pandemic (e.g., Litman et al., 2023).

In the attempted replication study, 15% of the participants reported behaviors like drinking or gargling bleach, using household cleaners on their bare skin, or inhaling bleach vapors to prevent COVID-19. After applying data quality techniques like those outlined above, the rate dropped to around 1%. However, when the behavior in question defies logic (like drinking bleach) it is critical to understand: is this 1% real or is it an artifact of unreliable participants who were not caught by the attention checks?

To answer that question, everyone who passed the attention checks and reported a dangerous behavior was asked to describe what happened in their own words. Among these participants, not a single person described drinking bleach to prevent COVID-19. What they did describe, however, were misunderstandings and misreadings of the question.

For example, when asked, "Did you drink or gargle soapy water to prevent COVID-19?" one participant, who passed all attention checks, answered 'yes.' When asked to describe what happened, they wrote: "My mother made me wash my mouth out with soap water because I was cursing, and I accidentally swallowed some."

This response reveals that the participant read the part of the question that mentioned drinking soapy water but apparently did not read the qualifier about whether this was done to prevent COVID-19. Without the open-ended question, the researchers would have incorrectly counted this as evidence that people were drinking soap as a COVID prevention strategy.

Another person who said 'yes' to the question, "Did you use household cleaner to clean or disinfect bare hands or skin to prevent COVID-19?" wrote: "I washed my hands with antibacterial soap." Regular handwashing with antibacterial soap, however, is not the dangerous

behavior the CDC was concerned about. Instead, it appears, this person thought antibacterial soap was an example of a “household cleaner.”

When another participant was asked, "Did you inhale the vapors of household cleaners like bleach to prevent COVID-19?" they answered ‘yes’ and for the open-ended questions they wrote: "I poured product on the floor and began to mop." This response reveals accidental exposure during routine cleaning rather than deliberate inhalation of bleach vapors to prevent COVID-19.

Overall, not a single participant described engaging in the dangerous cleaning practices that were the focus of the study. That means open-ended probes provide additional protection above and beyond what closed-ended attention checks provide. Between both approaches, it is possible to fully vet the responses in a sample even when it comes to a sensitive study that examines the prevalence of rare events. What began as a headline “Millions of Americans are drinking bleach” turned out to be the result of fraud, inattention, misreading, and misunderstandings—something researchers would never know without proper data quality measures.

Now that we have seen the value of assessing data quality through yea-saying questions, benchmarking items, and open-ended responses, let’s examine the various other kinds of attention checks researchers use to protect the quality of their data.

Types of Attention Checks

Module 11.2

Learn how low-quality data distorts descriptive, correlational, and experimental research

In the previous section, we examined how yea-saying questions can be used to remove inattentive and fraudulent participants from a study. Yea-saying questions are not, however, the only type of question researchers use to detect careless responses. In this section, we will review the main options available to behavioral scientists.

What Are Attention Check Questions?

Attention check questions are a tool of quality control in behavioral research. They do exactly what their name suggests: check if participants are reading the study materials and thinking about their answers.

Attention checks are used in studies to identify participants who might harm data quality. As we saw above, removing unreliable participants often yields different results because they provide systematically biased data (e.g., Litman et al., 2023; Oppenheimer et al., 2009). Hence, these checks protect the quality of research.

While attention checks have been around for decades (e.g., Petzel et al., 1973), they have become especially important in the era of online research (Arndt et al., 2022). Because researchers cannot see online participants or control the environment people take studies in, attention checks offer a way to ensure some level of participant engagement.

Types of Attention Checks

Researchers commonly use several kinds of attention check questions. Some directly measure attention, while others check for people's willingness to follow instructions or whether a study's manipulation worked as intended. Attention checks can also vary in whether they tap into attention, memory, or language skills. Below, we discuss various forms that attention checks can take, when each is used, and the potential tradeoffs of using different forms of attention checks.

Instructed Response Items

Instructed response items tell participants how to respond. They might, for example, say “select agree to show you’re paying attention.” (Gummer et al. 2021; Meade & Craig 2012).

Researchers often embed instructed response items within groups of similarly structured questions, as shown in Figure 11.8. Participants who are unengaged will likely miss these instructions.

How positive or negative do you feel about each of these terms?

	Extremely negative	Somewhat negative	Neither positive nor negative	Somewhat positive	Extremely positive
Fire Eater	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Fire Dancer	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Fire Performers	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Please select somewhat positive*	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Flame Throwers	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Pyromancer	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

→

Figure 11.8. A matrix of multiple-choice questions is often a good place to embed instructed response item attention check.

Longer versions of instructed response items measure whether participants have carefully read instructions or other typically long paragraphs (Oppenheimer et al., 2009). They often look like this:

"Most modern theories of decision making recognize that decisions do not take place in a vacuum. Individual preferences and knowledge, along with situational variables, can greatly influence the decision-making process. In order to facilitate our research on decision making we are interested in knowing certain factors about you, the decision maker. Specifically, we are interested in whether you actually take the time to read the directions. If not, then some of our manipulations that rely on change in the instructions will be ineffective. So in order to demonstrate that you have read the instructions please ignore the question text below and select the third answer from the bottom of the list as your answer.

Which of the following is your favorite hobby?

- Fishing
- Movies
- Gardening
- Reading instructions
- Walking
- Exercise
- Music
- Do not enjoy hobbies
- Other

If you read the entire question, then you likely noticed the instructions to select a specific answer, regardless of what your favorite hobby is. If, on the other hand, you skimmed the question or skipped over the instructions, then you, like the participants who fail these items, probably provided the incorrect answer.

Instructional manipulation checks work best when they look like other questions in the study and require participants to engage in behavior that requires carefully reading longer passages of text. Even then, however, there are often better ways of measuring whether people are reading and comprehending the study materials (e.g., Kane et al., 2023).

Creating artificial narratives for the sole purpose of examining whether people are reading the details of these passages can have negative consequences. People who correctly answer these items are likely to be reliable respondents. But among those who do not answer these items correctly, there may be good participants too. In other words, the longer the instructional item is the more likely it is to have false positives—people who are otherwise reliable respondents but

who may not have carefully read this particular paragraph. Instead of using long instructional items, we recommend using several shorter items dispersed through the survey.

In our experience, it is best to avoid long instructional items. In studies where it is important that participants carefully read and comprehend long passages, comprehension checks are a better option. Unlike instructional items that rely on arbitrary passages for the sole purpose of checking for attention, comprehension checks are tied to the stimulus materials of the actual study.

Comprehension Checks

Comprehension checks test whether participants understand the content of the study materials. They often focus on whether participants grasp what they are supposed to do by asking a few questions about the instructions or having participants explain the task in their own words. People who fail these items will probably perform poorly.

The instructions below are from a well-known economic task called the “Dictator game.” After reading instructions like these, participants may be asked to answer three closed-ended questions that follow, assessing their comprehension of the instructions.

This task is about dividing money between yourself and another person to whom you are randomly matched. You do not know this other person and you are unlikely to ever meet him/her.

You have been randomly assigned the role of the “allocator”. The other person is in the role of the “recipient”.

At the start of the task you will be endowed with \$100 and the recipient is endowed with \$0. You can decide how much of your \$100 to transfer to the recipient. You can choose any amount between \$0 and \$100.

After you make your proposal, the recipient's task is to decide whether to accept or reject your offer. Your payment and the responder's final payment depends on the responder's decision.

If the responder accepts, you and the recipient both receive the amount contained within your proposal. That is, you and the responder receive the amounts that you allocated to both of you.

If the responder rejects, however, neither you nor the responder receives any money; that is, both of you get \$0."

1. Which role have you been assigned to in the upcoming task? [Allocator, Recipient, Judge, Observer]
2. How much money will you be endowed with at the start of the task? [\$0, \$1, \$10, \$100]
3. What happens if the person you are paired with rejects the proposal? [We both receive \$0, We split the money 50/50, I make a second offer, The person makes a counterproposal]

Such comprehension checks are best used in cognitive psychological experiments or any study that does not contain a long list of survey questions. This is because studies that lack long lists of questions do not lend themselves to easily embedding yea-saying questions or instructional manipulation items into the survey flow. Instead, for experiments and other similar studies, adding several questions that check for people's comprehension of the study's instructions and stimulus materials is a good way to measure attention.

Manipulation Checks

Another way to use comprehension checks is to see whether respondents notice key elements of the experimental manipulation. As an example, imagine participants are asked to read a vignette that manipulates a person's occupation. Half of participants are told that a man works as an accountant and the other half are told he is a janitor. A manipulation check may ask participants to identify the man's occupation.

"What was John's occupation?"

- Accountant
- Janitor
- Entrepreneur
- School teacher
- Unemployed

Sometimes manipulation checks work at a group level rather than at the individual level. In such cases, manipulation checks examine whether the experimental manipulation produced a psychological difference between groups. Or it may assess how the information affects

participants psychologically. If the purpose of the manipulation was to signal different levels of socioeconomic status (high vs. low), a manipulation check might ask participants:

How prestigious was John's occupation?

1	2	3	4	5
Not very prestigious				Very Prestigious

Although these latter measures can increase a researchers' confidence that participants are paying attention at a group level, they do not help when looking at individual behavior. In addition, a lack of difference between groups does not necessarily mean participants weren't paying attention. It may simply mean the manipulation was not effective (see Hauser et al., 2018).

Attention Checks that Pull for Yea-Saying

As we saw in the previous module, yea-saying questions are a type of attention check that takes advantage of a behavior often exhibited by fraudulent and inattentive participants: their tendency to say "yes" and "agree" to most questions.

Yea-saying questions ask about behaviors or experiences that are either extremely rare or completely impossible. For example, asking participants if they have visited a small town with just a handful of residents, if they own items that do not exist, or if they have experienced statistically improbable events are all forms of yea-saying questions.

The effectiveness of these questions comes from how they tap into the strategic behavior of fraudulent participants. As you saw in Module 11.1, respondents working in click farms, not paying much attention, or otherwise attempting to game the survey know that agreeing with most statements helps them qualify for more studies. By presenting plausible-sounding but verifiably false questions, researchers can identify participants who are employing this strategy.

A key advantage of yea-saying questions is that they can appear natural within the survey. Rather than using obviously artificial checks like, "Have you ever had a fatal heart attack while watching TV?," researchers can create a subtle measure that blends into the survey's content. For

instance, asking about fictional products alongside real ones ("Which of these haircare brands have you used recently?") or inquiring about nonexistent venues ("Have you visited the Meridian Theater in Chicago during the past year?") makes these items a bit harder to detect.

Attention checks should aim to avoid the potential negative reactions that obvious attention checks can provoke, which range from participants feeling like their intelligence is being insulted to feeling the researchers are trying to trick them (e.g., Hauser & Schwarz, 2015; see Shamon and Berning, 2020).

Another advantage of yea-saying questions is their flexibility. Researchers can easily create new variations tailored to their specific research context, making them difficult for fraudulent participants to recognize and circumvent. While traditional attention checks often become recognizable to experienced survey-takers, yea-saying questions can be continually refreshed with new content while maintaining their effectiveness.

Open-Ended Checks

Open-ended questions ask participants to write a response in their own words. Often, these items work best when they are relevant to the study topic, but a general item our team sometimes uses asks participants to "Please describe the last thing you remember cooking and where you cooked it. Write at least one complete sentence."

Open-ended questions help researchers spot concerning behaviors like copying text from the internet or from AI, ignoring the question instructions, writing about an unrelated topic, providing generic responses common to click farms like "NICE" or "great product" in response to a question about breakfast, or using automated tools to fill in answers. These questions are especially good at catching people engaged in survey fraud, as these people often struggle to write natural-sounding responses. As you learned in Module 11.1, open-ended items can also ask participants to provide context for rare events or unusual behaviors that they endorsed in multiple choice items, checking the validity of their responses.

Issues With Attention Checks

Psychological Effects on Participants

Participants sometimes interpret attention checks as an attempt to trick them (Hauser & Schwarz, 2015; Silber et al., 2022). Thus, it should not be surprising to learn that attention checks can influence participants' behavior later in the study.

Some participants react to attention checks by deliberately failing the questions in an act of defiance (Silber et al., 2022). Others, tend to think more carefully about subsequent tasks, looking for hidden meaning (Hauser & Schwarz, 2015). And some see the questions as a challenge and pay more attention afterward (Kung et al., 2017; Shamon & Benning, 2020). These reactions show that attention checks are not neutral measurement tools. Instead, they are measures that can have a psychological effect on participants, just like everything else within a study (e.g., Hauser et al., 2018).

In our experience, people react negatively to attention checks when they are new to online research and do not understand why researchers care about data quality. Once people know the importance of these measures and that most studies include them, the measures become a mundane part of the survey process.

False Positives

One of the biggest risks with attention checks is false positives—incorrectly marking reliable participants as unreliable. As touched upon above, this can happen because participants misunderstand the question, inadvertently select the wrong answer, or make some other mistake. For example, someone who is asked, "Have you ever been to McMullen, Alabama?" might think of a similar-sounding town they really have visited. Or, someone who is asked about using fictional products might confuse the name with real brands.

Beyond individual questions, the risk of false positives increases when researchers set exclusion criteria that are too stringent. As you saw in Module 11.1, requiring participants to pass all attention checks can provide modest improvements in data quality but at the expense of removing many people who provided otherwise useable data.

Mitigating Negative Effects

How can researchers avoid the negative consequences outlined above? We believe the best way is to adopt the balanced approach outlined in Module 11.1: use multiple, relatively simple checks and require participants to fail at least two items before being classified as problematic.

Three characteristics make an attention check effective: 1) writing a clear question, 2) giving it one correct answer, and 3) ensuring the question measures attention rather than another construct.

Write Clear Questions

Attention checks should verify people are paying attention, not trick them. This means attention checks should be easy to pass *for people who are paying attention*. Effective questions begin with direct prompts that stand on their own. In several studies we have seen, researchers seem to assume the goal of attention checks is to trick participants. Questions like "If John's father's brother is married to Mary's sister, how are John and Mary related?" or "If today is Tuesday and yesterday was Monday, what day will it be three days after tomorrow?" assess much more than attention. They require participants to rely on logic and sequencing, working memory, and careful reading. Because these questions are confusing, they raise the risk of false positives.

Measure Attention, Not Other Constructs

Many researchers mistakenly use questions that assess memory, general knowledge, intelligence, or other cognitive constructs instead of attention. For example, in a study where participants read about two friends who have lunch to discuss work issues, asking "How many oranges did Steven have before trading with Josephine?" tests memory of peripheral details rather than attention to the study itself. Participants might have read but failed to recall minor details. Another way of putting this is, if the outcome of an attention question correlates with cognitive ability it is not a good attention question.

Consider another example that meets the elements we have discussed so far—clear prompt, single answer—but may, nevertheless, be challenging for participants: "How many times have you seen the 1997 film *Titanic* starring Brad Pitt?" Regardless of how many times someone has seen *Titanic*, the correct answer must be zero. That is because the movie starred Leonardo

DiCaprio, not Brad Pitt. But by assessing memory, the question relies on participant's prior knowledge, potentially tricking people working under time pressure. As a result, this question is a poor measure of attention.

A Strategy for Implementing Attention Checks

So, what should you do with the information above? What is the best way to protect your data quality?

Based on our experience with tens of thousands of researchers who have conducted studies with millions of participants, we recommend a strategy that balances simplicity with effectiveness, making it applicable across a range of research contexts.

For a standard 10–15-minute study, we recommend including 4 or 5 attention checks plus one open-ended item. As shown in Module 11.1, this combination has proven highly effective at identifying participants who provide low-quality data. The number can be adjusted based on your study's length and the sensitivity of your data, but this basic framework serves as a solid foundation for most research.

When using this approach, we suggest embedding your attention checks within groups of similar sounding items, so they don't clearly stand out. For example, in a study about social media usage you might include the question "Do you have an active Vines account?" among questions about social media platforms. Since Vines shut down in 2017, anyone answering "Yes" needs closer examination. This approach feels natural to participants while still effectively measuring inattention.

We also recommend ensuring your attention checks match the overall demands of your study. If your study asks people to read detailed text and notice small differences, then more complex checks might make sense. However, if your study mainly uses simple rating scales and multiple-choice questions, then simpler attention checks would be better.

When assessing people's performance, we recommend removing participants who fail more than one attention check or who provide an unusable open-ended answer. If the data you're gathering is descriptive in nature, you can require participants to pass all of your attention checks before they are included in the analyses.

Implementing Data Quality Measures in Your Research

Module 11.3

Explore a strategy to protect your online data collection

So far, you have learned about the effectiveness of attention checks and the variety of options available to researchers looking to assess data quality. Now, it's time to apply this knowledge.

In this project you will create a data quality strategy for a study you wish to run. In each step below, we will ask you to think about questions that will guide your plan for protecting data quality (see Figure 11.9). Your task is to think carefully about each question, select the best option based on your study, and then explain your thinking. Your completed assignment should include a detailed plan that is based on the information in this chapter and you should present your decisions in a clear, organized manner that demonstrates your understanding of the options for protecting data quality.

As you select items to assess data quality, you may draw upon the questions presented in this chapter, use AI to generate your own items, or visit a website called Survey Defense: <https://surveydefense.org>. At Survey Defense you can register for an account and then generate five randomly selected yea-saying questions from a large library of items (see Reevey et al. 2024). These yea-saying questions may not match the aims of your project. The goal of this project is to apply your understanding of the types of quality checks behavioral scientists use and how to embed them within a project.

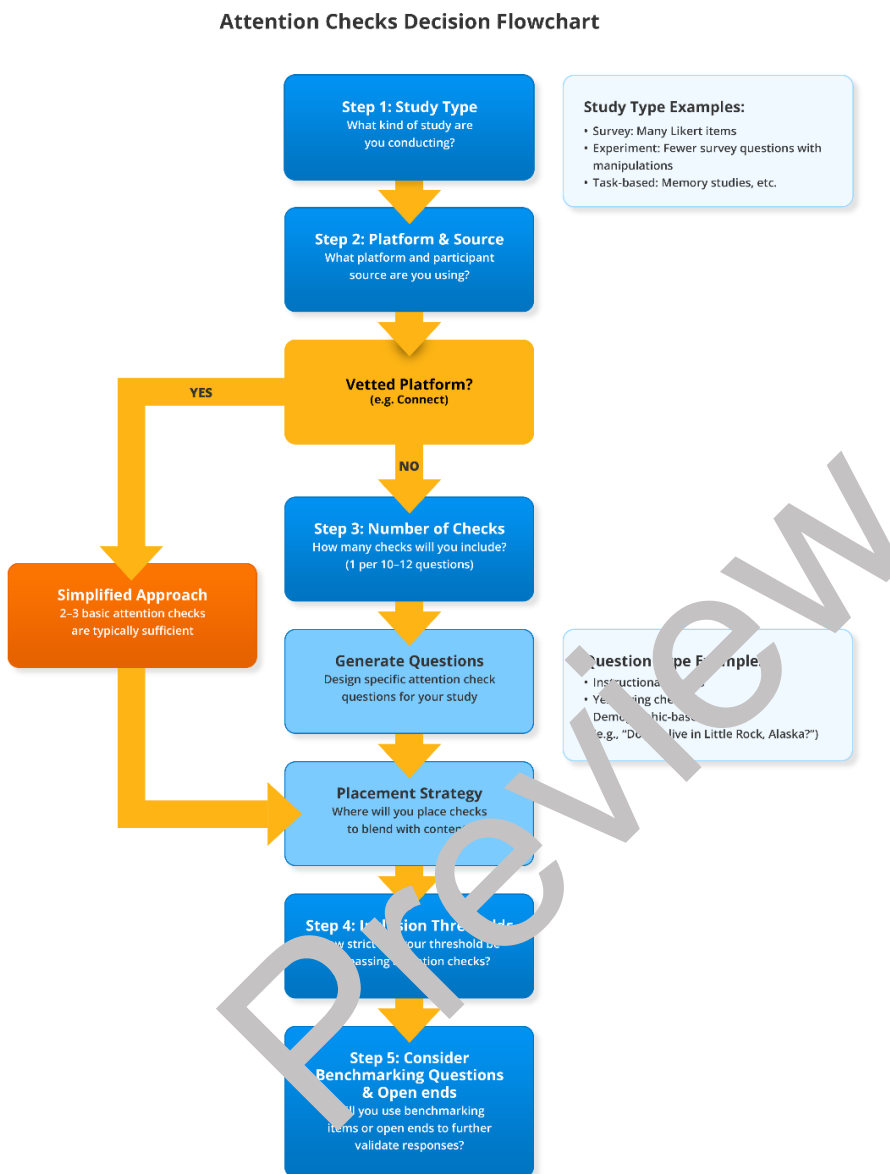


Figure 11.9. Questions that can guide your plan to protect data quality.

Step 1: What Kind of Study Are You Conducting?

Your plan to protect data quality will be shaped by the type of study you plan to conduct.

As you learned about in Chapter 10, descriptive research is more easily distorted by low quality data than correlational or experimental research. In addition, some studies such as experiments offer fewer opportunities to insert attention checks and yea-saying items than a survey that asks participants to complete several validated measures. Thus, your study will shape your approach to data quality.

In your assignment, begin by describing the type of research you are conducting. Is your study descriptive, correlational, or experimental? Next, describe what participants will do. Will they answer survey questions, or will they read an experimental vignette and respond to the situation? What opportunities does your study provide for measuring attention and data quality?

For example, let's say your survey is very short. Participants will be asked to read a short story and answer a handful of questions. With just a few dependent measures, you will not have the opportunity to embed more than a few attention checks. What you can do instead, however, is ask a few questions that ensure participants comprehended the instructions, the story, or the manipulation they experienced.

In your assignment, describe how the study enables or inhibits the inclusion of yea-saying questions and then discuss how important it is to ensure the data in your study are free of low quality responses. Remember that correlational and experimental studies can often tolerate more noise than descriptive research and exploratory studies can often tolerate more noise than studies that seek to confirm a theory or hypothesis in a long line of theoretical research.

Step 2: Where will you Find Participants?

In addition to the type of study you conduct, your source of participants affects your approach to data quality. As you learned in Chapters 9 and 10, some participant sources provide better data than others. In particular, crowdsourcing sites like Connect offer better data quality than market research panels because participants who sign up for a site like Connect are rigorously vetted before they are given access to studies (e.g., Hartman et al., 2023; Peer et al., 2023).

In your assignment, describe your source of participants and how much vetting or screening those participants will undergo before entering your study, if any. Explain how your sample source should affect your approach to data quality. If you are sampling from a source that

receives vetting before entering your survey, you might include fewer attention checks overall and structure your study to confirm people are paying attention and understood the big ideas like a manipulation they were asked to read or the kinds of tasks they are asked to do. If, on the other hand, your sample source has not been vetted or is known to have question marks around data quality, you will want a more rigorous approach to measuring that people are paying attention and providing honest answers.

Step 3: How Many Checks Will You Use?

Now, it's time to form your plan. How many attention checks will you use in your study? Why is this number appropriate? Your answers should draw upon both the type of research you are conducting and your sample source.

For example, if you are conducting a descriptive study with a source of participants known to have issues with data quality, such as a market research panel, you should adopt an aggressive approach to data quality. This might mean asking an attention check question every one or two minutes within your study. So, if the study lasts 10 minutes you might ask 4 or 5 attention checks. As the study gets longer, you can add more items to ensure participants are paying attention throughout.

If, on the other hand, you are drawing from a vetted source of participants, you might favor a more relaxed approach. In a 10 minute study, you might ask 2 or 3 attention checks.

As a rule of thumb, we recommend including approximately one attention check for every 10-12 survey questions. This means if the typical study lasts about 10 minutes and includes about 60 questions, perhaps five items will assess attention.

Generate Your Items

Once you know how many items you plan to include, you need to select each specific item. Remember, these should blend with the content of your study.

If you are conducting a survey or correlational study with many scale items, there are several types of checks you can incorporate into the study. You can easily add yea-saying questions, instructed response items, and at least one open-ended question. If you are conducting an experiment, on the other hand, you will have fewer opportunities to conspicuously add these

kinds of questions. Instead you might include comprehension checks that ensure people understood the instructions, a manipulation check that evaluates whether the manipulation provided a difference between experimental groups, and perhaps an item or two in the demographics section or elsewhere that assesses whether participants are paying attention. Create specific attention check questions appropriate for your study.

For your assignment, you should generate all the measures you plan to use. You may choose examples from the chapter, use the Survey Defense website, or even turn to AI to generate appropriate questions. For each question, explain why it's appropriate for the context of your study and what you intend to measure with it.

Decide Where You Will Place Each Item

Once you have your items, you must decide where they will go.

Describe in your assignment where you will place each attention check question and why it makes sense in that location. For example, if you are conducting a market research survey that asks participants about their experiences with different products, it may make sense to include a yea-saying question with products or services that do not exist early in the study where participants are used to seeing screening questions that ask about their experiences with different products. Or, if you are conducting an experiment that offers few opportunities for embedding attention checks throughout the study, you might opt to place a yea-saying question about vacationing in a small town such as New Rock, Indiana within the demographics section.

Overall, you should explain how you will space attention checks throughout your survey/experiment and how you will blend these checks with your study content so they don't stand out too much.

Step 4: What Criteria Will You Use to Exclude People?

The next step is to decide which criteria you will use to include and exclude participants. As with other decisions, this should be dictated by the type of study you are conducting.

Will you require participants to pass all attention checks? If so, why? If you allow participants to make a few errors, why is that decision defensible (e.g., 4 out of 5 correct)? Explain how your type of research and sample source informs this decision. Remember, if your study focuses on

frequency estimates or requires high representativeness, stricter thresholds (e.g., passing all checks) may be appropriate. For other studies, more lenient thresholds may be acceptable because they balance the risk of low quality data against false positives and the need to collect more participants in your study.

Step 5: Will you Include any Benchmarking items or Open Ends?

The final thing to consider in your data quality plan is whether you will include any benchmarking items or open-ended questions.

As demonstrated in the chapter, benchmarking items allow you to examine whether people who pass and fail your attention checks perform differently on items for which there is some objective assessment of truth (i.e., population base rates). The main determination about whether to include these items in your study or not is the survey length. Longer studies provide more room to surreptitiously include two or three of these items than shorter studies. But, as with other decisions, both the type of research you are conducting and the need to justify excluding participants should play a role in your decision. If you will add any of these items to your study, explain which rare activities you will ask about. You can adopt ones from this chapter or create your own.

Beyond benchmarking items, you should decide whether to include open-ended items in your study. As a reminder, if your study seeks to describe people's behavior or asks about rare events open-ended items can help verify participants' responses. In all other studies, these questions can be used in combination with attention checks to assess data quality. Explain if you plan to include open-ended items, what they will ask about, and how you will use them in evaluating data quality.

Summary

Throughout this chapter, we have explored how to identify the importance of data quality in behavioral research and the various methods researchers can use to identify and remove unreliable responses. As you have seen, low-quality data is a persistent challenge, but researchers have several tools to spot low quality responses and remove them from a dataset.

Among the effective approaches to protecting data quality you have learned about **yea-saying questions** are especially important. These questions leverage the tendency of fraudulent respondents to agree with most statements, allowing researchers to identify people providing unreliable answers. When used in combination with **benchmark validation questions** researchers can compare participants' response patterns to known population parameters, confirming the effectiveness of attention checks and validating the quality of the remaining data.

Beyond yea-saying questions, there are **various forms that other attention checks** can take including instructed response items, comprehension checks, and manipulation checks. Each of these question types has advantages in different research contexts.

By incorporating multiple attention checks, validating them with benchmark questions, and using open-ended items to verify unusual responses, researchers can significantly improve data quality. However, these strategies work best within a fit-for-purpose approach to data quality. Different research questions require different levels of quality control—studies examining associations between variables may tolerate some inattentive data, while research on rare behaviors demands more stringent screening and additional validation through open-ended responses.

When implementing these techniques, you should carefully consider potential trade-offs. While more stringent exclusion criteria bring data closer to ground truth, they also increase the risk of false positives—removing otherwise valid participants who may have made a single error. Finding the right balance requires understanding both your research objectives and the spectrum of data quality that exists in any sample.

As online research continues to evolve, maintaining high data quality standards remains essential. The tools and strategies outlined in this chapter provide a practical framework for researchers to ensure their findings accurately reflect reality rather than artifacts of inattention or fraud—ultimately advancing our collective pursuit of scientific knowledge.

Additional Readings

- Arndt, A. D., Ford, J. B., Babin, B. J., & Luong, V. (2022). Collecting samples from online services: How to use screeners to improve data quality. *International Journal of Research in Marketing*, 39(1), 117-133.

- Berinsky, A. J., Margolis, M. F., Sances, M. W., & Warshaw, C. (2021). Using screeners to measure respondent attention on self-administered surveys: Which items and how many?. *Political Science Research and Methods*, 9(2), 430-437.
- Kay, C. S., & Saucier, G. (2023). The Comprehensive Infrequency/Frequency Item Repository (CIFR): An online database of items for detecting careless/insufficient-effort responders in survey data. *Personality and Individual Differences*, 205, 112073.
- Litman, L., Rosen, Z., Hartman, R., Rosenzweig, C., Weinberger-Litman, S. L., Moss, A. J., & Robinson, J. (2023). Did people really drink bleach to prevent COVID-19? A guide for protecting survey data against problematic respondents. *Plos one*, 18(7), e0287837.
- Reavey, B., Bruggemann, P., Rosenzweig, C., & Litman, L. (2024). Sentry In-Survey: A tool for preventing survey fraud. [Manuscript under review. Contact authors for copy].

Chapter 12: Data Cleaning

Let's say you have designed your study, recruited participants, and watched the responses roll in. Now comes one of the most anticipated moments in research: diving into your data. But before you can explore those exciting findings, there is a crucial step that determines whether your results will be meaningful or misleading—screening your data for quality.

First, you need to confirm the data were properly collected. Did random assignment work, were there programming errors that prevented participants from answering questions, were all items properly recorded on consistent scales, did any stimuli or media get in properly displayed?

Next, you should ensure participants answered all (or nearly all) of the questions in the study. People with large amounts of missing data may need excluded from your analyses, and participants with small to moderate amounts of missing data may require action to replace their missing scores.³ While examining the completeness of your data, you should ensure each participant has only one row of data, and if you find duplicate entries, decide which row of data to keep (usually the first).

Finally, you should examine the quality of the responses. How did people perform on attention checks, did anyone provide patterned responses, are there outliers or other signs you can use to identify participants who may harm data quality? If so, you want to prevent these people from negatively influencing your analyses. Only after this methodical cleaning process can you create your analysis variables, check their reliability, and begin testing your hypotheses.

This process is not just academic housekeeping—it is essential for research integrity. Unscreened data can lead to misleading or entirely false conclusions, as you saw in Chapter 10. The percentage of questionable responses across online sources often ranges from 20-40%, making data screening a critical skill for any researcher working with online samples.

³ There are many ways to handle missing data. Some of them, like imputation of missing scores, are somewhat advanced topics that we do not cover here. Readers interested in learning more can consult Donders et al., 2006 or Enders (2022).

In Module 12.1, you will learn the fundamental techniques for screening data that work for most online research projects. We will show you how to download your data, evaluate attention checks by computing pass/fail scores, and assess open-ended responses for signs of fraud or inattention. You will learn step-by-step procedures for identifying which participants to exclude and how to make these decisions systematically.

In Module 12.2, we will explore advanced screening methods. You will discover techniques like individual consistency measures, response pattern analysis, timing data, and multivariate outlier detection. While these methods can be powerful, we will explain their limitations and why simpler approaches are often more practical for most studies.

Finally, in Module 12.3, you will apply these screening techniques to a real dataset from a marketing study seeking to develop a new snack. This hands-on experience will show you how poor data quality can lead to costly real-world mistakes and give you practice making the kinds of decisions you will face in your own research. You will see firsthand how to screen for data quality.

By mastering these screening techniques, you will ensure your hard work produces reliable insights rather than misleading conclusions. The credibility of your research depends upon this scrutiny—so let's learn to do it right.

Chapter 12 Outline: Data Cleaning

Module 12.1: The Fundamentals of Data Screening



Reading

Walk through the basics of screening and cleaning online data

- Learn about the purpose of data cleaning and how to approach it
- Examine evidence for the effectiveness of yea-saying questions
- Learn about tradeoffs in the sensitivity of these items and the value of open-ended questions

Module 12.2: Advanced Techniques of Data Screening



Reading

Learn about statistical measures of data quality and their limitations

- Explore the methods of assessing data quality that go beyond attention checks
- Learn about the strengths and limitations of statistical approaches to measuring quality

Module 12.3: Guided Application: Cleaning Data from a Product Development Study



Laboratory

Apply your knowledge to cleaning a real dataset

- Work with a real dataset to complete the data cleaning process
- Develop knowledge and skills you can apply to your own datasets

The Fundamentals of Data Screening

Module 12.1

Walk through the basics of screening and cleaning online data

The Purpose of Data Screening

Why must you invest time cleaning and screening data before answering your research questions? Why can't you get right to the fun stuff?

The answer is that unscreened data can lead to misleading or entirely false conclusions as we saw in Chapter 10. Screening data ensures the integrity and validity of research. We suggest you think of data screening as detective work. Your task is to gather evidence about each participant's engagement and compliance with study procedures. Like in the legal system, participants deserve the presumption of good faith participation until evidence suggests otherwise. When you identify signs of inattention, minimal effort, or failure to follow instructions, you must determine whether sufficient evidence exists to exclude their data from your analysis. The credibility of your research depends upon your scrutiny.

Removing Versus Rejecting Participants

The goal of screening data is to find participants who should be removed from your analyses. Ideally, you establish the criteria for excluding people before collecting data and document it in a pre-registration. Doing so prevents you from exercising what are called researcher degrees of freedom—judgments that may not be intentionally biased but nevertheless increase the odds of finding what you hope to find (Simmons et al., 2011).

When you find evidence that a participant failed to engage properly with your study, the appropriate response is to remove their data. When this process is pre-registered, transparent, and thoroughly documented, excluding low-quality data is not only acceptable, it is necessary. Including poor-quality data creates greater ethical concerns: wasted resources if true effects are obscured, or worse, publishing misleading findings that misrepresent human behavior.

A different and more consequential decision is whether to reject a participant's submission entirely. Unlike simply excluding data from analysis, rejection impacts the participant—they do

not receive payment and their reputation on the platform usually suffers. Some institutional review boards prohibit researchers from imposing these consequences, but this position is not universal.

We recommend reserving rejections for cases of clear fraud and excluding all other questionable data from analyses. This approach addresses the most serious threat to research integrity while acknowledging the uncertainty in assessing attention and engagement. Some platforms, including Connect, allow researchers to "flag" concerning participants without outright rejection. As participants accumulate flags and other negative feedback, their access to studies decreases, creating a system that incentivizes high quality.

Evaluating Attention Checks

If you followed the suggestions from Chapter 11, your survey contains some attention checks and open-ended items that can be used to evaluate data quality. In the space below, we demonstrate how to evaluate these items. Then, we describe some advanced methods of evaluating data quality, highlighting both when these advanced methods are appropriate and why our recommended approach is a simpler strategy fit for a wider array of online research projects.

Downloading Data

If you completed the assignment in Part 1, you know how to download survey data. You simply log in to your survey platform—Qualtrics, Engage, Redcap, etc—and look for a “Data” or “Data and Analysis” tab. The video we created for the descriptive research assignment in Chapter 3 shows how to download a Qualtrics data file, save the file to your computer, and open it in SPSS. You can find the video here: [link to video](#).

Once you find the “Data & Analysis” page in Qualtrics, choose “Export & Import” and then “Export Data” (Figure 12.1). You will be given several options for your file format. You can choose SPSS if that is your statistics program or select a CSV or Excel file to open the data in another statistical software. Once the data is downloaded, open the file and save it to your device.

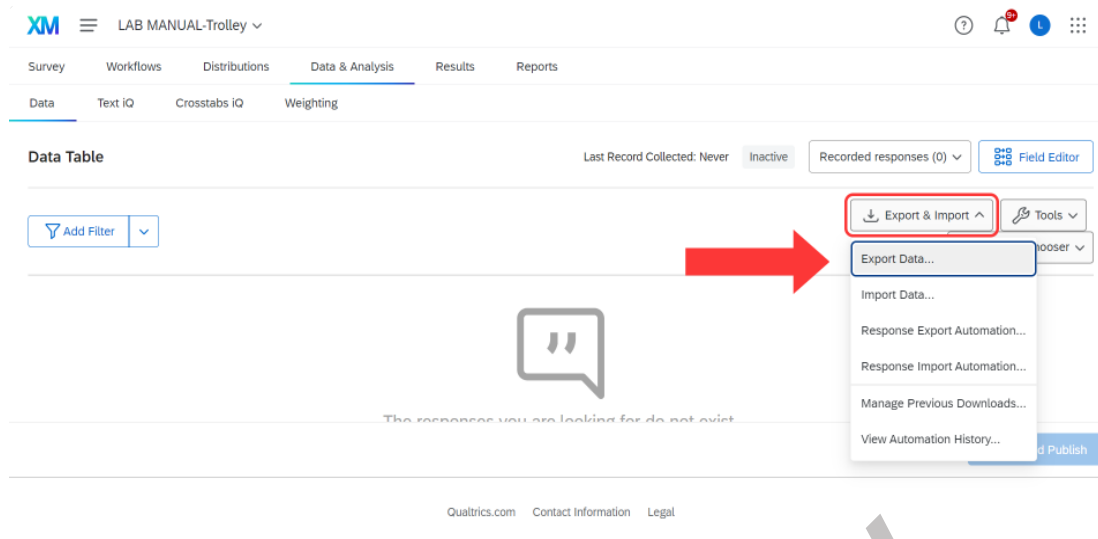


Figure 12.1. Under “Export Data” there are several options for downloading your data.

Computing an Attention Check Score

With your data downloaded, one of the first things you want to do is create a score that indicates how many attention checks each participant failed. People who failed too many items, should be excluded from the analyses.

Recoding Variables

Tallying up how many checks each participant failed requires recoding the questions into a pass or fail format. You might, for example, recode all the answers to the question “Are you currently a member of the LPAKE group?” (1 = yes, 2 = no) into a new variable named “LPAKE_pass” with a 0 = pass, 1 = fail format.

To recode a variable in SPSS:

1. Select "Transform" → "Recode into Different Variables"
2. Move your attention check variable into the “Input Variable -> Output Variable box” (Figure 12.2)
3. Name your new output variable (e.g., "LPAKE_pass")
4. Click "Old and New Values" to specify the coding scheme

While the LPAKE example above tells how to recode a question with a "Yes" or "No" response scale, imagine a 5-item scale where participants were directed to select "5 – Strongly agree." Using the dialog box, you would recode the old value of 5 into a new value of 0 (indicating a passing answer). Then, you could recode all other values into a 1 (indicating a failing answer). After you have recoded one attention check, repeat the process for all the other attention checks in your study.

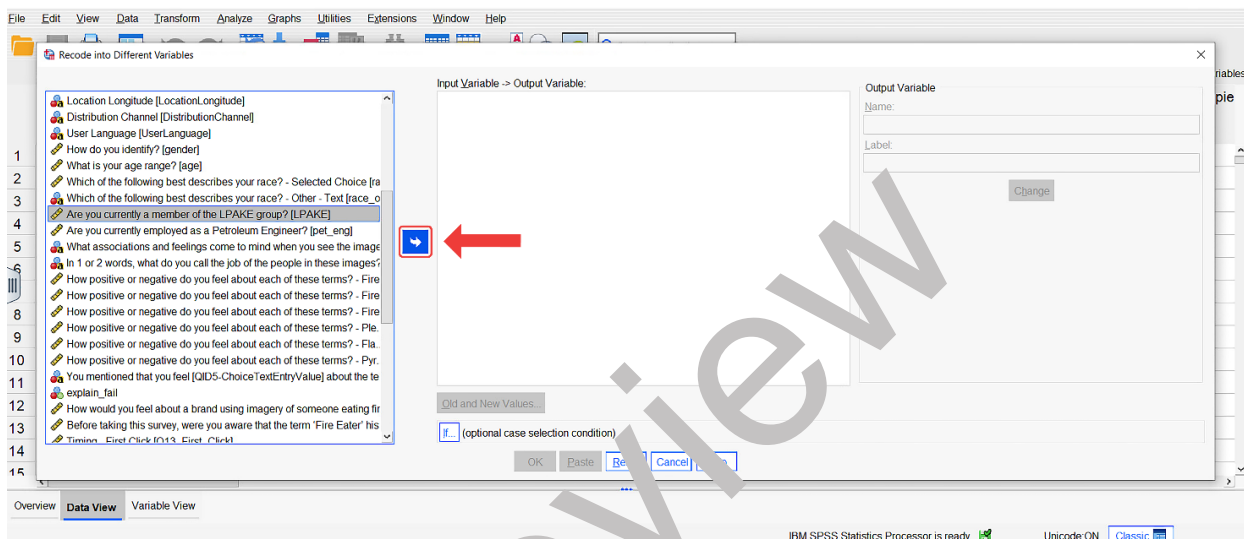


Figure 12.2. Once you have found the variable in the list, select it. Then, move it to the right using the arrow.

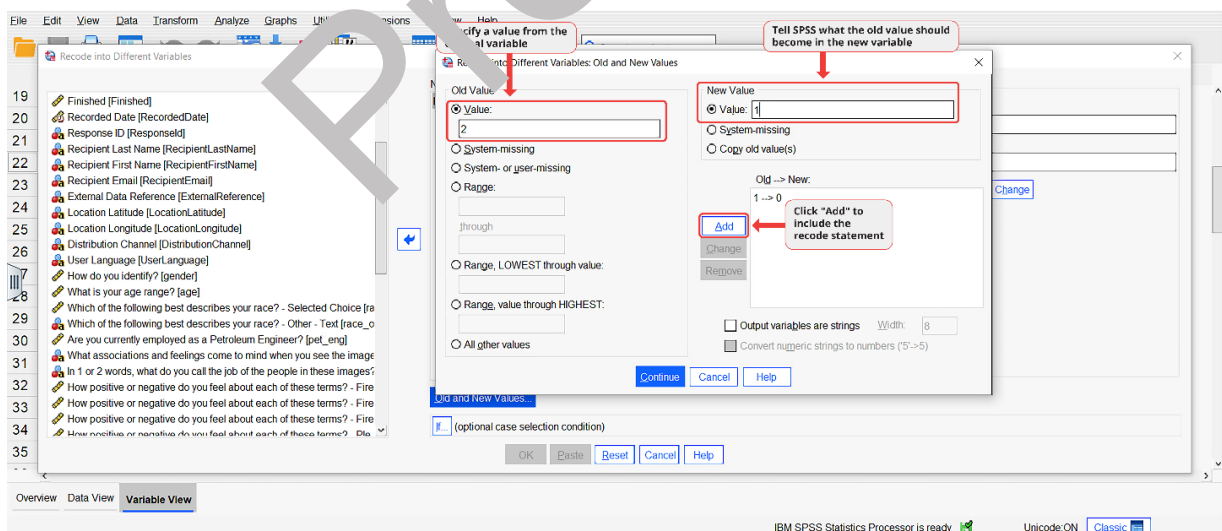


Figure 12.3. In the recode box, you can enter the old value, the new value, and then select "Add."

Computing an Overall Score

Once your items are recoded, you can compute an overall score that tells how many checks each participant failed. To do that, you will use SPSS's "Compute" function.

The "Compute" function allows you to create new variables based on a wide variety of mathematical operations. In this case, you will sum the number of attention checks each person failed. To do so:

Select "Transform" → "Compute Variable."

Type the name of your new variable into the "Target Variable" box. You might, for example, type "attn_total" (Figure 12.4).

In the 'Numeric Expression' box you will write the formula for the new variable. For instance, to sum three attention check questions named LPAKE_pass, petroleum_pass, and newrock_pass, you would enter:

Sum (LPAKE_pass, petroleum_pass, newrock_pass)

Finally, select Ok to execute the action or 'Paste' to add the command to a syntax file. Then check the new variable in your data file.

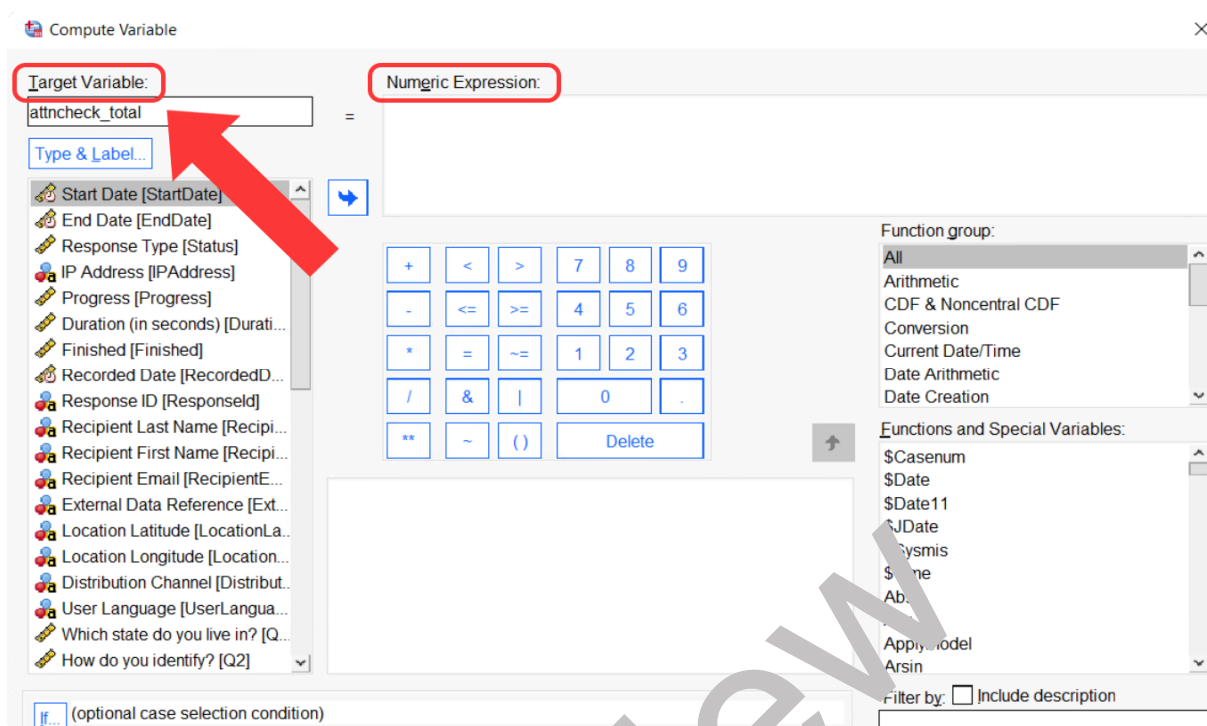


Figure 12.4. To name your new variable, type something into the “Target Variable” box. The “Numeric Expression” box is where you enter the formula for calculating the variable.

Tallying Pass and Fail Rates

With the overall score computed, you are ready to examine the percentage of participants who passed and failed the items and each person’s individual performance. Start by running a frequency or descriptive statistics report in SPSS.

1. Select “Analyze” → “Descriptive statistics” → “Frequencies.”
2. Find the variable(s) you wish to examine (e.g., LPAKE_pass). Move each item into the “Variable” box on the right and select Ok to run the analysis.

The Output from your analysis will appear similar to the data in Figure 12.5. There, you can see most participants passed the LPAKE question. Out of 1,075 participants in this example, 1,062 passed (98.8%). Just 13 participants failed the question.

→ Frequencies

[DataSet1] Z:\Connect P DQ Screening\Aug_Connect+Screening.sav

Statistics

Did the person pass the LPAKE question?

N	Valid	1075
	Missing	15

Did the person pass the LPAKE question?

	Frequency	Percent	Valid Percent	Cumulative percent
PASS	1062	97.4	98.8	98.8
FAIL	13	1.2	1.2	100.0
Total	1075	98.6	100.0	
Missing System	15	1.4		
Total	1090	100.0		

Figure 12.5. A frequency output table that shows the percentage of participants who passed and failed the LPAKE question.

From here, you can examine the pass rate for each item individually and the “attn_total” variable. People who fail two attention checks should be excluded from further analyses and may warrant a rejection. Meanwhile, people who failed one item warrant further scrutiny by looking at open-ended answers.

Evaluating Open-Ended Responses

There is no quick way to evaluate people’s open-ended answers. But, some answers are so reliably linked to fraud that it is worth the time to read people’s responses. In addition, open-ended items sometimes reveal quality issues that do not appear in attention checks or other “closed-ended” measures. Let’s look at how to evaluate open-ended answers.

Sorting the File

The first thing we recommend doing when examining open-ended responses is to sort the file in alphabetical order. Sorting allows you to easily spot duplicate responses or other patterns in what people have written.

To sort the file, navigate to the column with the open-ended question in the “Data View.” Right click on the column and select “Sort Ascending.” The answers will appear in alphabetical order.

After sorting the file, create a new variable to indicate if each participant passed or failed the open-ended item. Create this variable by right clicking next to the column you sorted and select “Insert Variable.” A new variable will be added to the file. You can name this variable something like “open_pass.” Then, you are ready to grade each response.

Grading Responses

Grading each response requires you to read it and judge whether it’s acceptable or not. We recommend scoring all passing answers with a 0, all failing answers with a 1, and any response you are unsure about with a 2. This allows you to easily sort the responses later.

As you grade, keep the information from Chapters 11 and 12 in mind. While most answers will be short and simple statements, some will be unrelated to the question prompt (such as “good”), will not follow the instructions (e.g., writing a single word when the question asked for a sentence), will appear copy and pasted from the web (e.g., “Fire-eaters were southern political ideologues who had uncompromising demands and played an important part of driving the nation”), or will contain such poor grammar that it is likely that the person is not a native speaker of the language the survey is in (e.g., “sometimes it makes accident”). In each of these cases, you should mark the response as a failure and exclude the participant from further analyses.

Excluding Participants

After evaluating your measures of quality, you must determine which participants to exclude from further analyses. We recommend excluding people who either failed multiple attention checks or provided poor-quality open-ended answers.

To implement these exclusions, create a variable named "exclude" or "drop" and assign a value of '1' to all participants you intend to remove from your analyses. Later, you can use the options in your stats package to exclude these people from analyses.

While participants who fail multiple attention checks typically provide poor open-ended answers as well, the reverse pattern is not always true. Some people who pass attention checks perform poorly on open-ended items because they lack the language proficiency to write coherent sentences. This is one reason why it is valuable to include open-ended questions in your screening process.

Sometimes the evidence from attention checks and open-ended responses will present a mixed picture. The appropriate course of action in these situations depends on your research objectives. Descriptive research requires vigilance against low-quality data because unreliable respondents can easily distort the research findings as detailed in Chapter 7. Correlational and experimental designs are more resilient to noise and some degree of bias.

Overall, the complexity of the decisions around data quality highlight the importance of preregistering your screening plan. When faced with ambiguous cases, consider analyzing your data both with and without questionable participants, while transparently reporting the decisions in your research.

Once you have decided on which respondents to exclude from analysis, it is worth comparing them to the included participants on benchmarking questions (see Chapter 11). Unreliable participants should provide noticeably higher levels of yea-saying or unrealistic values on benchmarking questions compared to those who are included, providing a measure of confidence that the right people were excluded from analyses.

In some situations, you might want to screen your data more closely than the techniques described above. The advanced techniques described next might help, but as we will see, they can be harder to implement.

Advanced Techniques of Data Screening

Module 12.2

Learn about statistical measures of data quality and their limitations

After learning about the attention checks and open-ended screening methods above, students often wonder: "Isn't there something more sophisticated? Aren't there screening methods that are more subtle, grounded in statistics?" The answer is 'yes.' There are numerous statistical methods for identifying low-quality responses in survey data (e.g., DeSimone et al., 2015). However, these approaches come with significant limitations that make them less practical than the approach outlined above.

While advanced techniques of data screening are hard for participants to circumvent, they also require more effort on the part of researchers to implement. They are also not always practical. In this section, we introduce some of the advanced methods for screening data, provide a conceptual understanding of how they are used, and highlight their strengths and limitations. For most studies, the approach described in Module 12.1 will be sufficient and easier to implement.

What Advanced Methods Do

The fundamental appeal of advanced screening methods lies in their covert nature. These measures use information that participants do not realize is being monitored to assess response quality. Unlike attention checks that participants can clearly identify, advanced methods work behind the scenes. And they analyze patterns in how people respond throughout the entire survey. The core assumption is that effortful, attentive responding should produce certain patterns in timing, consistency, and response choices, while careless or random responding will violate these patterns in ways a researcher can detect.

Advanced screening methods generally fall into three categories. First, consistency-based measures examine whether participants provide similar responses to items that should logically receive similar answers, or different responses to items that should receive different answers. Measures in this category include synonym-antonym pairs that identify similar or opposite items within a survey, and personal consistency measures that examine whether individual participants respond reliably within each scale they complete (DeSimone et al., 2015). It also includes

measures like the Squared Discrepancy Score, which examines the consistency of people's responses to scale items (e.g., Litman et al., 2017).

Second, timing-based measures assume that thoughtful responding requires a minimum amount of time and flags participants who complete surveys or individual items too quickly to have read and considered the content.

Third, pattern-based measures look for aberrant response patterns such as selecting the same response option repeatedly (straight-lining) or producing response combinations that are statistically unusual compared to the rest of the sample (Mahalanobis distance).

The appeal of these approaches is clear. Using measures participants cannot easily evade, researchers can catch people who are not paying attention. Furthermore, by adopting statistical approaches to identify who provides reliable and unreliable data in a survey, these advanced methods promise to catch different types of problematic responding in a manner that feels more objective than examining people's performance on attention check questions.

Why Advanced Methods are Hard to Implement

While advanced screening methods appear rigorous, they prove much harder to implement than many researchers initially expect. In fact, the practical challenges often make them impractical for most studies.

The first major hurdle is that these methods typically require long surveys with many items per construct to function effectively. Synonym and antonym pair approaches need sufficient items to reliably identify pairs with high correlations, which usually means having multiple items measuring similar concepts spread throughout the survey. Personal consistency measures also require multiple scales with numerous items each to produce stable reliability estimates. This creates a fundamental tension for researchers who want or need to conduct short studies. The brevity that makes online studies fast and affordable undermines the advanced screening methods that could assess quality.

Even with appropriately long surveys, however, there is no guarantee that reliable synonym or antonym pairs will emerge in any specific dataset. This means researchers often face a difficult choice: either invest extensive time before data collection identifying and pretesting potential

item pairs or wait until after data collection to identify pairs statistically. The first approach requires a substantial upfront investment of time that may be wasted if the planned pairs do not perform as expected in the sample. The second approach means the researcher will not know if the method is feasible until after the data are collected.

Perhaps more surprisingly, for methods that appear statistically rigorous many advanced screening techniques require researchers to make subjective judgments about things like cutoff scores. Because measures of reliability or consistency can vary from measure to measure, there may be few established benchmarks for any particular measure. This means many decisions are made without clear empirical guidance about what constitutes reasonable standards.

Finally, the technical expertise required to implement several advanced measures means that implementing just one of these measures might require a researcher to learn new statistical procedures, write code to calculate screening indices, troubleshoot computational issues, and validate their implementation—all before even knowing whether the approach will work with their data. This makes these methods particularly inappropriate for beginning students who are still mastering fundamental research skills and may lack the statistical background needed to implement and troubleshoot complex screening procedures.

The cumulative effect of these challenges means that advanced screening methods often consume far more time and resources than anticipated, with uncertain benefits that may not justify the investment. Many researchers discover that straightforward attention checks and open-ended screening questions provide a good assessment of data quality with far less complexity.

When Advanced Methods Might Be Worth Considering

When is the right time to consider advanced measures?

They can be worthwhile in large-scale studies that include multiple scales or several items and where researchers have the time to develop expertise with the methods. An example of this kind of research might be scale construction, when researchers test many items, often in a long survey, to find out which ones work best for assessing different psychological constructs. High-stakes research where data quality is paramount is another instance in which these measures might make sense. Similarly, established research programs where investigators can develop and

validate screening approaches across multiple studies may find the initial time investment worthwhile.

For most student research projects and many professional studies, the combination of well-designed attention checks and open-ended questions provides a more practical approach to screening. The sophistication of advanced methods does not automatically make them superior—often, simpler approaches that you can implement well are more valuable than complex methods that introduce their own complications.

The key insight is that good data screening should match your study's specific needs, timeline, and expertise rather than defaulting to whichever method appears most statistically sophisticated. Sometimes the most rigorous approach is the one you can execute with the time and resources you have.

A Final Word: Avoiding False Positives

When considering how to exclude data, it is useful to adopt a multifaceted approach. Instead of excluding people based on a single measure, it is better to accumulate multiple pieces of evidence. This method addresses the problem of false positives (dropping someone who provided a sincere effort).

For instance, in situations we have encountered, some researchers will dismiss participants by evaluating each measure individually (e.g., excluding five people for timing data, excluding 10 people for failing an attention item, and excluding 5 people for open-ended answers, etc). Such an approach can lead to unnecessary exclusions and the loss of valuable data (as the previous module demonstrated, where you draw the line on excluding participants matters). Instead, we recommend accumulating multiple pieces of evidence that a participant provided poor data before excluding them from your dataset. Most often this will mean failing multiple attention check measures or providing an open-ended response that indicates fraud.

While it is sometimes necessary to exclude several participants, these instances are rare when sampling from well-vetted researcher centric platforms like Connect. When sampling from platforms like Connect, researchers often only need to remove small portions of their data. If you are excluding more than 30% of participants, there may be an issue with your measure, materials, or screening process.

Guided Application: Cleaning Data from a Product Development Study

Module 12.3

Apply your knowledge to cleaning a real dataset

After collecting data, you have the chance to identify quality issues and remove bad participants from your dataset. You will have the chance to practice this with an example project. Then you will be ready to apply the techniques to your data.

Working through this example will help you develop the following knowledge and skills:

1. You will learn how to begin analyzing any dataset you encounter by first becoming acquainted with the measures and examining indicators of quality. You will then learn how to handle participants who perform poorly on measures of quality.
2. You will learn to distinguish between participants who should be removed from a dataset and those who should be rejected from the study.
3. You will learn how to filter data so that only high-quality responses are included in your analyses.
4. By screening data for quality, you will develop the knowledge to better plan data quality checks, depending on the kind of research you are conducting.

The activities in this section give you experience with an important part of the research process: cleaning data. When poor quality data goes undetected, it can ruin your research. So, let's learn how to properly identify and remove poor responses from your data.

Study Background: Fire Eaters, Spicy Snacks, and Brand Reputation

Imagine this.

You work for a large multinational food corporation. Your team is creating a marketing campaign to coincide with the launch of a new line of spicy snacks. You design a survey and recruit participants online.

In the survey, you measure what people think of marketing the new snack under the concept of 'fire eaters.' In addition to asking people for their impression of fire eaters, you ask if people are

aware that the term used to refer to pro-Slavery Southern Democrats within the United States. After informing people of this historical connection, you ask again whether they approve of a brand using modern-day ‘fire eaters’ in an ad.

You find three things in your data.

1. People feel positively toward ‘fire eaters.’
2. Many people (35%) know the history of the term ‘fire eaters.’
3. Even after people are informed about the history of ‘fire eaters,’ few disapprove of a brand using modern-day fire eaters in an ad.

With these results, your team develops a marketing campaign that costs hundreds of thousands of dollars. So far, this is exactly how research is supposed to work.

But there’s a catch. Unbeknownst to you, the data from your survey is low quality because the platform used for sampling has poor procedures for vetting participants. No one on the team thought to include quality measures beforehand, meaning the dataset is full of fraud. In addition, given what you know about survey fraud, you are reasonably sure that unreliable respondents have systematically skewed the responses toward positive answers, which means the study *overestimates* how positively people feel toward ‘fire eaters’ and *underestimates* how many people think the campaign is a bad idea. As the data analyst for this project, your job is to screen the data and remove the low-quality responses.

If this situation seems far-fetched, it is not. This is the basic situation that researchers at Kellogg’s encountered before they began working with CloudResearch. You can learn about their story at <https://www.cloudresearch.com/fire/> and you can dig into the dataset below.

Assignment: Clean the Kellogg’s Survey Data

1. Examine data quality in the “Fire Eaters” dataset.

On the *Research in the Cloud* OSF page you will find a folder labeled Chapter 10 – Cleaning and Screening. Within this folder you will find the fire eaters dataset. The first thing you should do is acquaint yourself with the survey. The accompanying .pdf shows the measures from the study.

After you look at the measures, try assessing quality on your own. Look at how many people passed the attention checks using descriptive statistics. Then, examine the open-ended responses. Mark which participants you consider to be reliable and unreliable. After attempting this on your own, check your work using the instructional video. How closely do your results match the example?

2. Report the results of your cleaning.

In a document, write up the results of your screening. Report how many people passed and failed each attention check question. How many people were excluded for bad open-ended answers and what kinds of reasons were they excluded for? At the end of your screening, how many participants were removed, and perhaps most critically, does leaving those people in the dataset change the takeaway from the main measures—measures like what percentage of people know about the term fire eaters and whether people think the campaign is a good idea.

Summary

Data screening is a critical step that determines whether your research findings will be meaningful or misleading. This chapter demonstrated how to screen data so you can avoid misleading conclusions.

Typically, the data screening process involves three steps. First, a researcher downloads the data and ensures the file is complete, things worked as expected in the study, and the data are ready to analyze. Second, the researcher evaluates the quality of each participants' responses. Finally, the researcher decides who to exclude from analyses.

We demonstrated a straightforward approach to data screening that is suitable for most online research, and we gave you an opportunity to practice that approach with a real dataset from a marketing study. We also introduced more advanced measures and described both their strengths and weaknesses for online studies. Regardless of how exactly you screen your data, it is important to remember the goal of data screening: removing unreliable responses that can mislead your research findings. In the next chapter, we describe how to design survey studies so that they are easy for participants to navigate, improving data quality.

Additional Readings

- Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2014). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American journal of political science*, 58(3), 739-753.
- Hartman, R., Moss, A. J., Rabinowitz, I., Bahn, N., Rosenzweig, C., Robinson, J., & Litman, L. (2023). Do you know the Woolly Bully? Testing era-based knowledge to verify participant age online. *Behavior research methods*, 55(7), 3313-3325.
- Kung, F. Y., Kwok, N., & Brown, D. J. (2018). Are attention check questions a threat to scale validity?. *Applied Psychology*, 67(2), 264-283.
- Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological methods*, 17(3), 437.
- Rivera, E. D., Wilkowski, B. M., Moss, A. J., Rosenzweig, C., & Litman, L. (2022). Assessing the efficacy of a participant-vetting procedure to improve data-quality on Amazon's Mechanical Turk. *Methodology*, 18(2), 126-143.

Chapter 13: Best Practices for Designing a Study

This book is about online research but, for a moment, we want you to think about *offline* research.

Imagine you need to conduct a study, but modern technology is not available. Participants cannot use computers to answer questions or complete tasks that measure reaction time. You cannot use tablets or mobile devices for games, website testing, or tracking behavior. Your study is limited to what you can observe or measure while in the same space as participants. What can you do?

In the past, behavioral scientists used these situations to observe people's behavior, record their conversations, and measure just about every psychological construct that can be measured with paper questionnaires. It may surprise you, but this requires a lot of work.

First, a researcher must create questions (and ideally go through the steps of validation outlined in Chapter 4). Then, they must format those questions into a document and make copies. If the study includes different versions or varies the order of questions, the researcher must prepare multiple versions and give each participant the correct one. After participants complete the questions, their responses must be scanned into a computer, which is a slow process prone to error.

Now, compare this with online surveys. When creating digital surveys, researchers can choose from different question types and control how they appear. The digital format allows for interactive, flexible, and personalized materials. If the study requires participants to watch a video, view an image, listen to audio, or read a scenario, the researcher can easily present these materials. They can also change the order of questions based on people's previous answers, guide participants down different paths in the study, and remind participants of their previous responses before asking new questions. The survey tool automatically records people's responses and can even collect information about the participant's device or browser. Simply put, online surveys offer more possibilities with less effort than paper ones.

In this chapter, we will explore how to design online studies. If you read Part I or if you have used survey tools before, the information at the start of this Chapter may be familiar to you. For

example, we start in Module 13.1 by describing two foundational parts of online survey platforms: survey elements and question blocks. But regardless of your experience with survey platforms this chapter will also extend your knowledge by describing how to best program online surveys and summarizing what research has to say on how participants respond to different aspects of an online study in Module 13.2.

By the end of this chapter, you will understand the potential and flexibility of online survey platforms; the different types of questions you can present to participants; essential skills for programming surveys and customizing online studies; and how to make design choices that align with your research goals.

Preview

Chapter 13 Outline: Best Practices for Designing a Study

Module 13.1: Understanding Survey Platforms and Their Features



Reading

Learn how survey platforms work as digital tools for creating and distributing online research studies

- Explore different survey elements including interactive question types like drag-and-drop, heat maps, and sliders
- Understand how to organize surveys using question blocks with randomization and branching logic

Module 13.2: Best Practices for Survey Design



Reading

Apply best practices for survey design including when to use matrix questions versus standalone items and the optimal length for response scales

- Learn how to make decisions about question formatting such as scale direction, vertical versus horizontal alignment, and open versus closed-ended items
- Explore how to implement advanced survey features like response requirements, timed text, timers, and embedded data to enhance data quality

Understanding Survey Platforms and Their Features

Module 13.1

Learn the basics of survey platforms and how they operate

What is a Survey Platform?

If you have never used a survey platform, think of it as a digital toolbox that allows users—everyone from researchers at prestigious institutions to students in research methods courses—to create customized questionnaires, distribute projects, and collect data efficiently.

Survey platforms have transformed data collection in the behavioral sciences by making powerful tools accessible to anyone who can use a point-and-click interface—no technical knowledge or advanced programming skills required. Beyond creating projects, many platforms include basic data analysis tools that help users identify patterns through simple visualizations and descriptive statistics. Despite their sophisticated features, survey platforms serve a straightforward purpose: they allow researchers to ask questions and gather responses.

For simplicity, we refer to all materials presented to participants within an online study—instructions, consent forms, individual questions, experimental manipulations, etc.—as "survey elements." These elements form the basic building blocks of a study. To show you the range of these elements, we will walk through a study about consumer reactions to a new product. This example will demonstrate how some survey elements recreate traditional offline questions while others introduce new ways to collect data.

Another important concept within online surveys involves what are known as blocks. Blocks are how researchers organize elements. Blocks give online researchers the flexibility to do things that are either difficult or impossible offline. For example, blocks make it possible to randomize content and create branching logic—features that customize each participant's experience in the survey. To show how powerful blocks can be, we will revisit the study we introduced in Chapter 4 examining the relationship between anxiety and depression.

Finding a Survey Platform

Survey platforms are abundant, yet when it comes to selecting the right platform for your project, you should consider what participants need to do and how many responses you plan to collect. At minimum, your chosen platform should provide tools to design your study, distribute the project, and securely store participant data.

Your first step should be checking whether your institution has an existing subscription. Many colleges and universities subscribe to Qualtrics or similar services. If you do not have access through a subscription, Qualtrics offers a free version that allows up to 500 responses. For simple projects that require more than 500 responses, consider Google Forms. While not as powerful as Qualtrics, Google Forms allows unlimited responses and works well for straightforward projects. Other common platforms include [Alchemer](#), [RedCap](#), [Gorilla](#), [SurveyMonkey](#), and [QuestionPro](#). CloudResearch's Engage is also available for use with this book.

Survey Elements

Survey elements typically vary throughout a project. At the beginning, researchers often present study information and consent forms. In the middle, they might embed videos, scenarios, images, or audio recordings and collect participants' responses to various questions. At the end, researchers frequently ask demographic questions, provide debriefing information or direct participants to helpful resources. While the beginning and ending portions are usually informational, the middle section contains the core research activities.

To show you the range of survey elements available in online research, we are going to describe a project that the authors conducted for a company called SpongeBath. While we talk about SpongeBath, this is not an advertisement—we have no financial interest in the company.

SpongeBath sells a product that cleans dirty sponges. As their marketing says: "The sponge is often the dirtiest object in your house!" The product works by soaking sponges in a cleaning solution, making them more sanitary when you need to use them.

When SpongeBath approached us, they were just starting out. Their main goal was to understand how to market their product. They asked for a market segmentation study to learn who might buy SpongeBath, what issues those people cared about, and how to best communicate the product's

value. In our study, we used many different question types that demonstrate the range of survey elements available online.

Instructions and Images

We began the SpongeBath study by clearly describing the task. We told participants we wanted their advice about which messages should appear on the product label. We also emphasized the importance of honest, thoughtful responses and thanked people for their time. This information appeared in what is typically called a "text" or "instruction" survey element.

After the introduction, we showed participants pictures of the SpongeBath product and cleaning solution (Figure 13.1). We mentioned that we would later ask about the label. The ability to present images is a basic feature within online survey platforms.



Figure 13.1. Images are a type of survey element researchers can use to present information to participants.

Video, Audio, and Other Media

After participants examined the product, we asked them to watch [a two-minute video](#) showing how Spongebath works.

To present the video, we embedded a URL into the survey platform. Most survey platforms allow researchers to insert media from the internet or upload files from their own device. This easy incorporation of files gives online research projects tremendous flexibility, as videos and images can become part of experimental manipulations.

Multiple Choice Questions

Next, we presented participants with multiple choice questions. We asked how interested they were in purchasing the product (extremely, somewhat, slightly, or not at all interested), and how much they would be willing to pay (\$0, \$0-\$5, \$5-\$10, \$10-\$15, etc.). Throughout the study, we used multiple-choice questions to explore different topics.

Open-Ended Items

After the multiple-choice questions, we included an open-ended item: “Please tell us what you think about this product in your own words. Feel free to share any thoughts you have such as questions, concerns, or general comments about the product.” Participants typed their answers in a text box.

Open-ended items are valuable when researchers want participants to express their own thoughts without being limited by predefined answer options. These questions are especially useful for exploring new topics or gathering fresh ideas. Chapter 8 describes how to collect more in-depth open-ended responses using AI-based survey platforms.

Ranking Items

Another question type we used was a ranking question (Figure 13.2). Here, participants ranked a set of options according to their preferences. We included these items because SpongeBath's creators already knew which messages they wanted to communicate—they just needed to learn which ones would be most effective.

Below are 5 pieces of information regarding sponges. Please drag and drop these messages, and rank order them in terms of how effective you think they would be in raising people's interest in the product.

- Even after one use your sponge is dirty.
- Every time you use your sponge it gets dirty.
- What is the dirtiest item in your home? The kitchen sponge.
- Sponges are dirty after every use.
- The average kitchen sponge is 200,000 times dirtier than your toilet seat.

>>

Survey Completion

0% 100%

Figure 13.2. This question asks participants to rank messages in terms of perceived effectiveness.

Why ask participants to rank items instead of using several multiple-choice questions to measure the same preferences?

Multiple choice questions (sometimes called rating questions) typically ask participants to evaluate something on a scale. They are useful for measuring the intensity of people's feelings, perceptions, or behaviors. Ranking items, on the other hand, show the relative priority of each answer option and reveal how several items compare to each other. Ranking questions can also be more engaging for participants, improving their overall experience.

Despite these strengths, ranking items do not show how much participants prefer one option over another. They also become unwieldy when the list of items is long. The choice between ranking or multiple-choice questions often depends on whether you want to measure the strength of attitudes or establish a preference order.

Drag and Drop

Drag-and-drop items offer another interactive way to gather online data. They work well for ranking questions, categorization tasks, and matching exercises. Participants in our study completed the drag and drop question shown in Figure 13.3.

This question helped us understand which messages participants found most compelling about SpongeBath. They dragged potential messages into either "Include on package" or "Do not include on package" categories.

Drag and drop questions make surveys more engaging. They provide a visual way for participants to organize items, creating a digital version of sorting tasks that traditionally required face-to-face studies.

We will now ask you some questions about the product that you just saw.

We would like to put various messages on the package to increase people's interest in the product. Below are a list of messages that we are currently considering to be put on the package.

We are specifically interested in what you think would be the most effective COMBINATION of messages to put on the package. Please drag 5 messages that you think should all be included on the package in order to maximally increase people's interest in the product. Please drag these to the 'included on package' box.

Also please drag five messages that you think would be least effective to the 'Don't include on Package' box.

Items	Include on package
The all natural, citric acid refill cleans your sponge without harmful chemicals	
Antimicrobial ingredients give you a clean sponge every time you need one	
SpongeBath is eco-friendly	Do not include on package
No more smelly sponges	
Save money by extending the life of your sponge by three times	
No harsh chemicals	
Non-toxic cleaning solution	
SpongeBath effectively cleans and kills odor-causing bacteria	

Figure 13.3. A drag and drop question allowing participants to indicate which messages they considered effective for SpongeBath product packaging.

Heat Maps

A unique online question type is the heat map. This asks participants to look at an image and select areas they like, dislike, or find interesting. As participants interact with the image, the system records data and creates output like what appears in Figure 13.4.



Figure 13.4. Heat maps are unique to the online environment. They allow researchers to see which areas of an image or webpage participants click, like, or focus upon.

The heat map shows where participants clicked during the task. In our study, people selected portions of the image they did not like. As you can see from the red areas in Figure 13.4, certain parts of the label were unappealing to participants.

Heat maps offer an interactive experience for participants and reveal interesting patterns for researchers, though they are not relevant to all projects.

Sliders

The final question type we will highlight is the sliding scale. These have an offline equivalent called the visual analog scale (VAS).

Sliders feature an interactive bar with a movable pointer that allows participants to visually show their position on a scale. While sliders can be designed with or without marked intervals, they generally offer a more engaging alternative to traditional scales. We did not include sliders in the SpongeBath survey, but Figure 13.5 shows what these questions might have looked like.



Figure 13.5. A slider scale can be used to assess participant's level of confidence or the percentage

Sliders capture a participant's position on a continuum. Like many question types, sliders can be harder to use on mobile devices than on desktop computers. With fine-grained scales (like 0 to 100), some device settings make it difficult for participants to select the exact point they want.

Consider which devices your participants might use or perhaps limit participation to desktop computers when using sliders.

Summary of Survey Elements

Online surveys give researchers many ways to present materials to participants—instructions, questions, and multimedia content. Individually, these survey elements are the smallest building blocks within an online project. Together, they can be combined in ways that engage and interest participants.

Yet beyond individual elements, researchers must consider how to arrange the pieces of a survey. The next section shows how they do this with question blocks.

Organizing Elements with Blocks

A question block is a group of survey elements within an online project. Figure 13.6 shows a block containing three elements: first, a short set of instructions; second, a multiple-choice question about anxiety; and third, a question that records how long participants spend on the page.

Two important features of this block are worth noting. First, some elements are separated by page breaks, which allow researchers to display content on different screens. Second, the questions in the block relate to each other. When building a project, researchers typically group questions on similar topics or from the same instrument within a single block. This approach makes organizational sense and gives researchers control over when and how those elements appear to participants.

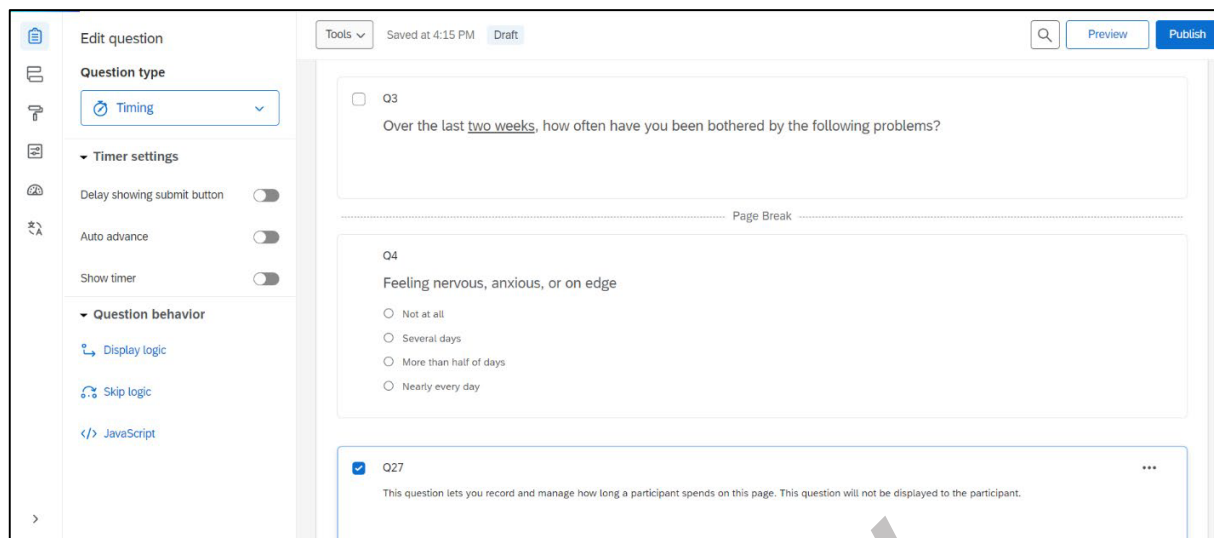


Figure 13.6. An example block with three elements: instructions, a multiple-choice question, and a page-timing item.

To demonstrate the benefits of organizing items in blocks, let's revisit the correlational study from Chapters 4-6. In that study, we examined the relationship between anxiety and depression. Anxiety was measured using the GAD-7 (see Table 3.1 for items) and depression using the PHQ-9. Below, we will show how questions can be arranged in blocks and highlight several important considerations.

Arranging Blocks

There are two ways that arranging blocks can be useful within an online survey. First, you can randomly determine the order in which blocks appear. Second, you can direct participants down different paths based on their previous answers. The value of both options becomes clear when we examine what they allow researchers to accomplish.

Randomization

Most survey platforms contain a randomization tool (also called a randomizer). This tool allows researchers to present blocks in random order. Without this feature, it is impossible to conduct an experiment or counterbalance the measures within a study.

Between-Subjects Randomization. The most common use of the randomization tool is for between-subjects experiments. Remember from Chapter 7 that each participant in a between-subjects experiment is randomly assigned to one experimental condition, regardless of how many

conditions exist in the study. By randomly assigning participants to conditions, the researcher ensures all groups are equivalent except for the independent variable they experience.

Using the randomization tool for a between-subjects experiment is straightforward. As shown in Figure 13.7 (top panel), researchers simply enter the survey flow area of the platform. Within the survey flow, they add a randomizer and place each block that should be randomly presented under it. Using the available options, researchers can ensure each participant sees only one block and that blocks are evenly distributed across participants.

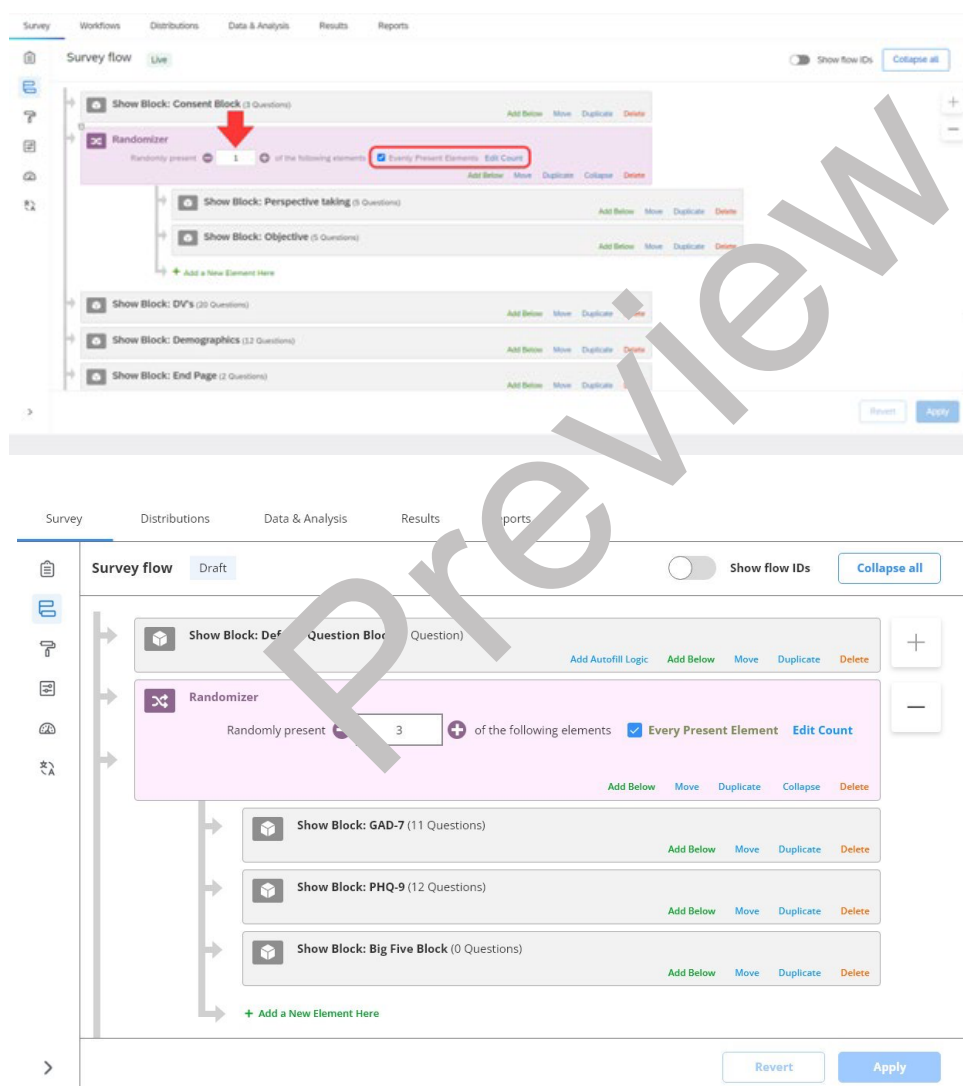


Figure 13.7. Top panel: Between-subjects randomization assigns one block per participant. Bottom panel: Within-subjects randomization presents all blocks to each participant in a randomized order.

Within-Subjects Randomization: Counterbalancing and Repeated Measures.

Randomization is useful beyond between-subjects experiments. For instance, in a correlational study measuring anxiety and depression, researchers might worry that answering questions about depression could influence people's responses to questions about anxiety (or vice versa). If so, they could counterbalance (randomly determine) the order of the questionnaires. When measures appear in a random order to different participants, order effects cannot explain any association between the two constructs.

You can see how within-subjects randomization differs from between-subjects randomization by comparing the top and bottom panels of Figure 13.7. In the between-subjects design, participants see just one block beneath the randomizer. In the within-subjects design, they see all blocks, but in a random order determined for each participant.

Beyond correlational studies, within-subjects randomization is useful for repeated measures experiments. In these experiments, participants experience every level of the independent variable. As with correlational studies, the only way to prevent carryover effects is to randomize the order of conditions.

Branching and Skip Logic

Another way to tailor the survey experience is through branching and skip logic. When used properly, these tools allow researchers to present participants with relevant questions based on their previous answers. This customization is not necessary for all research projects, but it saves time when used effectively.

Branching not only maintains participant engagement but also improves data accuracy. For example, in a survey about treatment experiences for various medical conditions, branch logic can be used to skip treatment questions for conditions a participant has not experienced.

Programming a survey branch can occur either within the survey flow or within a block. After adding a branch to the flow, the researcher must specify the logic it will follow (Figure 13.8). For example, a researcher might specify that if a participant indicates they have never been diagnosed with a particular condition, the survey should skip to the next condition, the next block, or wherever the researcher wants the participant to go.

The screenshot displays a survey flow editor interface. At the top, there are navigation tabs: Survey, Workflows, Distributions, Data & Analysis, Results, and Reports. The current view is 'Survey flow' in 'Draft' mode. A 'Show flow IDs' toggle is turned off, and a 'Collapse all' button is visible. The flow consists of several blocks:

- Show Block: Opening Block (1 Question) with 'Add Below', 'Move', 'Duplicate', and 'Delete' options.
- Show Block: Flat Earth Block (6 Questions) with 'Add Below', 'Move', 'Duplicate', and 'Delete' options.
- Show Block: Big Five Block (2 Questions) with 'Add Below', 'Move', 'Duplicate', and 'Delete' options.
- A conditional branch block with a blue header and a yellow warning: 'This branch will not be triggered until you Add a Condition'. It has 'Move', 'Duplicate', 'Options', 'Collapse', and 'Delete' options. Below it is a '+ Add a New Element Here' prompt.
- Show Block: Conspiracy Beliefs Block (1 Question) with 'Add Below', 'Move', 'Duplicate', and 'Delete' options.
- Show Block: Demographics (1 Question) with 'Add Below', 'Move', 'Duplicate', and 'Delete' options.

At the bottom right, there are 'Revert' and 'Apply' buttons. A large 'Preview' watermark is overlaid diagonally across the lower half of the image.

Figure 13.8. Adding branches in the survey flow.

Best Practices for Survey Design

Module 13.2

Explore what research has to say about how to best program a survey

When designing your study, you will face several important decisions about how to present survey elements to participants, regardless of your research topic.

Consider a study that measures personality with the Big Five Personality Inventory (BFI; John et al., 2008). The BFI contains 44 items that form a single instrument. It is best to present the BFI in a single block, but within that block a researcher could choose to place all the items in a single group (formally known as a matrix question; see Figure 13.9), as individual items on separate pages, or in some other configuration. Which is best?

Or consider another issue. Notice in Figure 13.9 the personality items use a five-point scale. Is five the optimal number? How do we know participants shouldn't be given 6, 7, or 11 answer options?

These questions bridge survey design with survey methodology—the study of how to effectively develop, implement, and analyze survey data. The answers matter because they affect both the experience of participants and data quality. In fact, research shows that participants cite structural and visual design issues as one of their biggest frustrations with online surveys (Fowler et al., 2023), highlighting the importance of getting these elements right.

Below, we discuss methodological considerations in online surveys. We outline current best practices for presenting different question and answer options, combining empirical evidence with our practical experience.

I see myself as someone who...

	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
1. Is talkative	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
2. Tends to find fault with others	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
3. Does a thorough job	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
4. Is depressed, blue	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
5. Is original, comes up with new ideas	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
6. Is reserved	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
7. Is helpful and unselfish with others	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
8. Can be somewhat careless	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 13.9. A questionnaire with eight items from the Big Five Personality measure.

Matrices or Standalone Questions?

When your project includes multiple questions, as nearly all projects do, you will need to decide whether to program the questions as a matrix or as standalone items. For example, in the project examining the association between anxiety and depression from Chapter 4, our anxiety measure had 7 items, and our depression measure had 9 items. To these measures, we added several questions about demographic variables and measures about people's behavior in other contexts related to anxiety and depression such as sleep. This meant that we had dozens of survey elements that required careful presentation.

Matrix questions share common response options and allow researchers to group similar questions together. For that reason, matrix questions are easier and faster for participants to answer than multiple items spread across several pages (e.g., Couper et al., 2001; Tourangeau et al., 2004). Presenting all items on one page may also improve their reliability (e.g., Couper et al., 2001). For all these reasons, matrix questions are appealing.

However, large matrices can cause participants to get tired or bored. Bored participants are more likely to drop out of the study, answer randomly, skip questions, or provide uniform answers—the behavior known as "straight-lining" (Iglesias et al., 2001; Toepoel et al., 2009). Matrix questions can also display poorly on mobile devices (Revilla et al., 2017). Each of these issues can reduce data quality, so it is important to use matrix questions judiciously.

Standalone questions work better when each item requires careful consideration or has distinct response options (Debell et al., 2021). Single-item questions increase the likelihood that participants focus on one item at a time, and although single items may take longer for participants to complete, they reduce the risk of patterned responses (Roßmann et al., 2018). Single items also generally display better on mobile devices.

Research shows that participants prefer standalone questions, despite the longer completion time (Thorndike et al., 2009; Toepoel et al., 2009). Large matrices tend to correlate with lower response rates and more dissatisfaction among participants (Debell et al., 2021; Toepoel et al., 2009). This dissatisfaction comes from what some participants describe as "bubble hell" (Fowler et al., 2023). With many small "bubbles" displayed on the same page, participants can easily miss one or lose track of the answer options as they scroll down the page (Figure 13.9). Some survey features address these issues, but a large page of text is generally harder for participants to navigate than a single item.

Instead of large matrix questions, participants prefer approximately three items per page (Grandmont et al., 2010). They also prefer questions with a small to moderate number of response options (e.g., five or fewer), which prevents visual clutter (e.g., Fowler et al., 2023; Liu & Cernat, 2018). Both findings deserve consideration when designing your project, especially when matrices are necessary. For example, presenting all items from a lengthy measure like the

BFI one at a time might be too tedious. Instead, a series of small matrix questions would be appropriate.

When matrix questions are necessary, there are ways to reduce the burden on participants. Some platforms let researchers repeat the answer option labels after every few rows in a matrix question, making it easier for participants to see which options correspond to each bubble. Some platforms also offer dynamic question features, such as graying out answered rows or highlighting missing answers, which can decrease missing data (Couper et al., 2013; Galesic et al., 2007). However, other interactive tools, such as highlighting rows when the mouse hovers over them, may actually increase missing data (e.g., Kaczmirek, 2011). Other options such as static shading of alternate rows or using background colors and typefaces to distinguish items appear to have no effect (Couper et al., 2013; Kaczmirek, 2008). In most cases, one of the most helpful features is requiring a response for each question, which forces participants to provide an answer for items they might have accidentally skipped.

To apply this advice, consider a survey that measures anxiety and depression like the project in Chapter 4. The GAD-7 contains seven items measuring anxiety. Even though the measure is brief, there are several ways to present it. We could display all seven items on one page, as shown in Figure 13.10.

With this configuration, some participants may need to scroll to see all questions, depending on their device. The question text can also appear cramped, especially on mobile devices. Since this measure contains just seven items, breaking the matrix into separate questions on different pages makes sense. We could present between two and four items per page for optimal participant experience.

When questionnaires are long, deciding how many questions to present per page can be challenging. Our recommendation is that two to four items per page is ideal for most participants.

Over the last two weeks, how often have you been bothered by the following problems?

	Not at all	Several days	More than half the days	Nearly every day
Feeling nervous, anxious, or on edge	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Not being able to stop or control worrying	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Worrying too much about different things	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Trouble relaxing	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Being so restless that it is hard to sit still	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Becoming easily annoyed or irritable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Feeling afraid, as if something awful might happen	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

[→](#)

Figure 13.10. The seven questions from the GAD-7 presented in a single matrix.

How Many Response Options?

When selecting the number of answer options, it is important to balance participant engagement, mental effort, and the precision needed in responses. A number that consistently balances these needs is five (e.g., Adelson et al., 2010; Donellan & Rakhshani, 2023; Rakhshani et al., 2023; Simms et al., 2019).

Five-option answer scales give participants room to express themselves without being overwhelming. If your project requires more nuance, a 7-point scale might be appropriate, but research shows little improvement in data quality with scales longer than 6 options (e.g., Simms et al., 2019). In other words, longer scales generally won't improve your results.

Related to the number of response options is whether to use an odd or even number of choices. This question has received significant attention because Likert scales with an odd number of

options include a neutral middle point. Even when this middle option is not labeled, participants interpret it as a neutral position (Nadler et al., 2015; Kulas & Stachowski, 2009). In contrast, even-numbered scales have no middle point, which requires participants to lean one way or the other. Figure 13.11 illustrates this difference. The top scale forces people to either agree or disagree with the statement, while the bottom scale allows for a neutral response.

I know where to seek support for my mental health

Strongly disagree <input type="radio"/>	Somewhat disagree <input type="radio"/>	Somewhat agree <input type="radio"/>	Strongly agree <input type="radio"/>
--	--	---	---

I know where to seek support for my mental health

Strongly disagree <input type="radio"/>	Somewhat disagree <input type="radio"/>	Neither agree nor disagree <input type="radio"/>	Somewhat agree <input type="radio"/>	Strongly agree <input type="radio"/>
--	--	---	---	---

Figure 13.11. Response scales with and without a middle option. Scales without the middle option force participants to take a side.

Consider omitting the middle point when your survey addresses familiar topics, when you are concerned about social desirability bias (Johns, 2005), or when participants might be uninvolved with the subject (Weems & Onwuegbyzie, 2001). Removing the middle option encourages participants to take a position. However, including a middle point is better when participants might genuinely hold neutral opinions, such as when the topic is unfamiliar (Johns, 2005). Your survey's context will often help determine whether to include a neutral option.

The context should also guide your choice about the number of response options. If you are using a validated instrument like the GAD-7, this decision has been made for you. When creating

your own questions, prioritize simplicity and clarity to keep participants engaged. Aim for 4 to 7 response options per question (see Chapter 4).

Regardless of how many response options you choose, maintaining consistency across questions helps participants understand and respond to your survey. Whenever the response scale changes, highlight this change through instructions or a visual break in the survey, often created with a page break.

When deciding on the appropriate number of response options, always consider your research goals and the mental effort required from participants as they move through your project.

Direction of Response Options

After deciding how many response options to include, you will need to consider which direction your scale should run.

Traditionally, scales progress from positive to negative (e.g., agree to disagree). Many researchers believe this direction feels intuitive and easier for participants to understand. However, for that very reason, you might occasionally choose to run your scale from negative to positive (disagree to agree). This change can encourage more thoughtful engagement and reduce participants' tendency to acquiesce (e.g., Chyung et al., 2018; Garbarski et al., 2019).

Several factors influence how participants respond to questions. People tend to agree more with items presented early in a survey than items that appear later. They also agree more when answering questions with longer scales compared to shorter ones, and when responding to mixed questions rather than items that clearly measure one concept at a time (Yan et al., 2018).

Given these findings, it is best to use shorter answer scales and to measure one construct at a time rather than mixing different measures together. This approach helps maintain data quality regardless of scale direction.

You should also remember that cultural factors may affect how participants respond to your questions. Not everyone reads from left to right, and what is considered 'intuitive' varies by culture. That means you should consider your participants' cultural background when designing

your survey, especially when collecting data internationally—something online research makes extremely easy.

As with the number of response options, maintaining consistency in scale direction throughout your survey helps avoid confusion and improves data quality.

Vertical or Horizontal Answer Scales?

After determining the number and direction of your answer options, you must consider alignment—whether the scale is presented vertically or horizontally (Figure 13.12). In terms of data quality, research suggests the answer to this question does not matter much.

Considering the images and videos you have seen of the Earth (e.g., from space agencies), how would you describe its shape?

- Definitely round
- Likely round
- Unsure/Don't know
- Likely flat
- Definitely flat

Considering the images and videos you have seen of the Earth (e.g., from space agencies), how would you describe its shape?

- | | | | | |
|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| Definitely round | Likely round | Unsure/Don't know | Likely flat | Definitely flat |
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

Figure 13.12. The top panel shows a vertical scale, and the bottom panel shows a horizontal one.

Knowing that scale alignment has little impact on data quality (Chan, 1991; Friedman & Friedman, 1994; Hu 2020; Toepoel et al., 2009) means you are free to base this decision on

factors such as readability, response accuracy, and your survey platform's capabilities rather than concerns about data integrity.

Vertical alignment works better on mobile devices because it is easier to read on smaller screens. It also benefits surveys with lengthy questions or multiple-choice options by providing a clear view of all answer options, reducing mental effort for participants.

Horizontal alignment is more common on desktop computers and particularly suitable for matrix questions. Whichever format you choose, ensure all answer options are visible without requiring participants to scroll, as this could introduce response bias.

Understanding your target audience is important when choosing between vertical and horizontal alignment. The devices participants use vary significantly by recruitment source. In market research panels, approximately 50% of participants use mobile devices, with even higher rates in developing nations (Chandler et al., 2019; Moss et al., 2023). In contrast, less than 10% of participants on crowdsourcing platforms typically use mobile devices (Chandler et al., 2019).

As with other aspects of your survey design, maintaining consistency in alignment throughout your survey helps prevent confusion and supports participant engagement.

Open or Closed-Ended Items?

Given the many factors that can influence participant responses beyond the actual question content, you might wonder if open-ended questions—where participants freely express their thoughts in writing or audio—would be more effective.

Open-ended and closed-ended questions serve different purposes in online surveys. Open-ended questions are particularly valuable when exploring new topics, investigating areas where the range of potential responses is not fully known, or seeking detailed, qualitative insights from participants. These questions allow participants to respond in their own words, which can reveal unexpected insights and perspectives that researchers might not have anticipated.

While open-ended questions provide valuable data, they present certain challenges. The responses can be complex and time-consuming to analyze, often requiring thematic analysis and coding to identify patterns (for a comprehensive look at qualitative research, see Yin, 2016).

Additionally, since these questions demand more effort from participants, the quality of responses may vary considerably.

Closed-ended questions, conversely, are used to quantify opinions and behaviors. Their structured format facilitates straightforward comparisons between groups and across different time points. However, closed-ended questions constrain the range of responses available to participants.

Your choice between open-ended and closed-ended questions should align with your research objectives. Open-ended questions are used for exploratory research and capturing the full spectrum of human experiences. Closed-ended questions are used for descriptive or experimental research where specific hypotheses are being tested.

Many effective surveys combine both question types, leveraging the efficiency of closed-ended questions and the depth of open-ended ones. Chapter 8 explains how AI tools are enabling researchers to conduct qualitative and mixed methods research at the scale of quantitative research, and the Engage platform specifically lowers the barrier to conducting open-ended responses by allowing participants to either type or speak their response using voice to text technology.

Other Survey Options

In addition to the features above, survey platforms possess several tools that do not need to be a part of every project but make certain projects more flexible and dynamic than offline ones. In most cases, these tools offer functionality that is not possible without technology.

Response Requirements and Question Validation

Earlier, we recommended creating no more than five questions per page throughout the survey to improve presentation and reduce missing data.

Most online platforms offer another tool for making sure that participants do not accidentally skip questions: response requirements. There are two basic types: you can request that participants answer overlooked items, or you can force participants to respond.

In our experience, requesting a response works better than forcing one. The request function reminds people about unanswered questions but allows them to proceed if they choose not to answer. Forcing responses prevents participants from moving forward until they have answered everything on the page, which might lead to inaccurate data if participants feel compelled to answer questions they prefer to skip.

Beyond response requirements, many platforms allow you to specify formats for participant responses, known as question validation.

Question validation ensures participants provide answers in a consistent format or within a specific range, which simplifies data analysis. For example, in the SpongeBath survey, we asked participants how much they would pay for the product. To avoid bias, an open-ended question was ideal, allowing participants to answer without a suggestion from the researcher.

However, open-ended responses can be messy. Some participants might write “\$20,” others “20 bucks,” and some simply “20.” While these examples are easily converted to dollar values, other answers are more ambiguous. For instance, if asking about household income, an answer of “20” could mean \$20,000 per year, \$20 per hour, or something else.

Question validation prevents these issues by requiring answers to follow a specific format. While this makes data easier to analyze, overly restrictive requirements can frustrate participants. To avoid this, provide clear instructions and consider adding an optional comments section for clarifications. Always pilot test any custom validation added to your survey.

Piped Text

Survey platforms offer piped text (also called dynamic text) options that personalize the survey experience.

Piped text inserts content from various sources—such as previous responses—into subsequent questions or answer options. This works through a line of code that pulls information from other parts of the survey and displays it to participants.

You can use piped text to modify survey questions based on specific criteria like demographic information, responses to previous questions, or external data sources. This feature allows you to personalize experiences and contextualize questions and answer options. When used effectively, piped text transforms a standard survey into an engaging, personalized research tool.

Timers

Online platforms make it possible to time participants. The most basic timing measure records how long people take to complete the survey. Those who finish a long survey very quickly probably ignored instructions or rushed through questions. Conversely, those who take an unusually long time on a short survey were likely distracted or stepped away.

The value of timers extends beyond tracking completion time. Researchers can embed timers that either: (a) ensure participants spend a minimum amount of time on a page or (b) limit how long participants have before the survey automatically advances.

In one version of the SpongeBath study, we asked people to watch a short video. While we could not guarantee their attention, we increased the likelihood of people watching by preventing them from advancing until enough time had passed to view the entire video. In this way, timers can encourage engagement with study materials.

Timers can also create pressure. Some researchers study how cognitive processes differ when people have limited versus abundant time. In these studies, researchers might present a stimulus but give participants only a brief period to view it before automatically advancing the survey when time expires.

Embedded Data

Embedded data is information recorded about participants outside the traditional question and answer format. This might include:

- Information about the webpage participants visited before entering the survey (useful for marketing and consumer behavior research)
- Data about participants' location during the study
- The experimental condition a participant was assigned to
- A calculated value based on other answers within the survey

- Information captured from the participant’s URL

This last option helps researchers record each participant’s unique ID, linking participation records from a recruitment platform like Connect to survey responses. To understand how this works, imagine your survey’s URL is normally “<https://www.example.com>.” Many recruitment platforms add values to this URL, such as:

“<https://www.example.com?participantId=ABCD1435&hitId=35hgfsbdu&assignmentId=vbwufg>.”

Using embedded data, a researcher can capture the additional values for storage in the data file. To capture the participant ID, the researcher would enter “participantId” in the appropriate field within the survey platform (Figure 13.13). The platform will then record this information during each participant’s session.

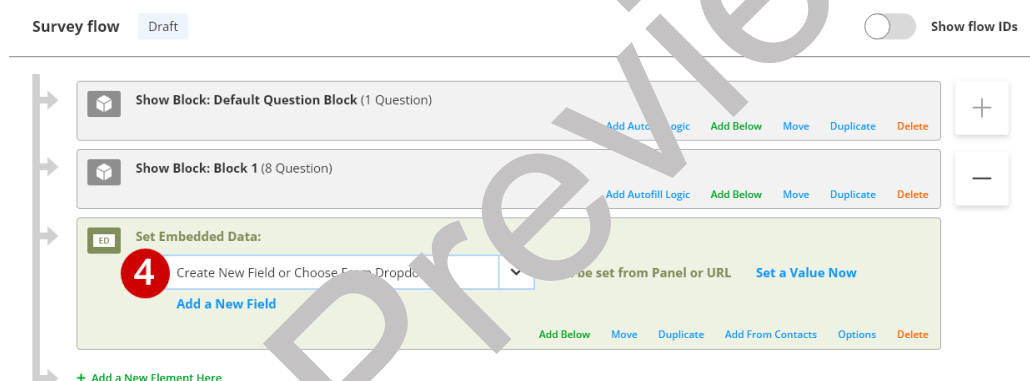


Figure 13.13. Embedded data allows the researcher to specify which values to capture from the participants’ browser, device, or survey behavior.

Progress Indicators

Online surveys can display progress bars showing how far participants have progressed through the survey. Are these bars helpful?

On one hand, progress indicators make studies more engaging and might reduce dropout by showing participants how much of the project remains. They may also motivate participants by giving them a sense of momentum as they move through the study. On the other hand, it is easy to envision how participants may feel discouraged if they are not progressing through the survey

quickly enough. If people get too discouraged, they may drop out rather than finish the study, or they may begin to rush lowering the quality of the data.

Progress bars do not significantly reduce drop off (Liu & Wronski, 2018; Villar et al., 2013) and have mixed effects on other aspects of data quality. While you may choose to include the progress indicator to inform participants about how much time they have left, you should not count on a progress indicator to deter participants from leaving the survey early.

End of Survey Options

Every survey must end, and there are multiple ways to program this conclusion. In some cases, a simple thank you message is sufficient.

When recruiting from paid sources, it is common to end with a redirect or completion code after debriefing participants. Both serve the same purpose: they confirm who completed the entire project. Note that they do not indicate data quality.

Specific instructions for programming redirects or completion codes are usually provided by the recruitment platform. Typically, you create an end survey element in your survey flow that directs participants who complete the project to a specific URL. Redirects are slightly easier for participants and more widely used, but completion codes work equally well.

Regardless of how your survey ends, include an option for participants to provide feedback about the survey itself. This feedback might clarify issues with participants' data or offer valuable suggestions for improving future surveys.

Estimating Survey Completion Time

Accurately estimating completion time is important for several reasons: it helps determine appropriate compensation, allows you to communicate honestly with participants (affecting their ability to plan participation), and prevents issues like dropout, rushing, or incomplete data that result from inaccurate time estimates.

Pilot testing provides the best estimate of completion times. Some recruitment platforms have built-in pilot systems that allow you to launch your project with a small number of participants

before opening it fully. Without this option, you might ask friends or colleagues to complete your project. For accurate estimates, these testers should be unfamiliar with your measures and should approach the project as actual participants would.

When pilot testing is not possible, estimate survey length by allowing about 10 seconds per question (Litman & Robinson, 2020). Multiplying your total question count by 10 provides a rough estimate of completion time in seconds.

Summary

Online survey platforms have become essential tools in modern behavioral science. They empower researchers to create dynamic projects, distribute surveys and experiments across the internet, and collect quality data from participants worldwide.

Designing effective and engaging online surveys requires understanding the tools discussed throughout this chapter—the various survey elements available to present to participants, how to select appropriate questions and organize them into blocks, how to optimize question presentation on screen, and when to implement advanced features such as randomization, skip logic, branching, page timers, and end-of-study options. With this foundation, you can evaluate the programming of the guided research projects presented earlier in this book. You should recognize how each study incorporates the principles outlined in this chapter to ensure high-quality data collection from a design perspective.

Despite careful planning, there may be times when the guidelines presented here are not suited to your specific project or participant pool. Human behavior is unpredictable, and research rarely proceeds exactly as anticipated. To address this uncertainty, pilot testing remains invaluable. The feedback gathered during testing can significantly improve your survey before full deployment. The next chapter will provide detailed instructions on conducting effective pilot tests and setting up your project with a participant recruitment platform to collect data.

Chapter 14: Best Practices for Setting up and Launching Online Studies

This chapter offers best practices for setting up and launching an online research study.

We will cover essential decisions such as how to name and describe your project to attract appropriate participants while avoiding selection bias, how to determine fair and effective compensation, how to use platform tools for targeting specific participant groups, and how to apply demographic quotas. We will also discuss strategies for improving representativeness, including how to demographically match the sample to the US census. We will then discuss how to best configure your study to capture necessary participant identifiers for tasks like longitudinal tracking or bonus payments, while still protecting anonymity. Finally, we will cover best practices for launching a study, including considerations for when is the best time to launch, and how to monitor data collection, work with participant feedback, and address any technical issues that may arise.

Then, we will examine more complex research designs and strategies required for their successful execution. We will focus on best practices for conducting longitudinal research, with a particular emphasis on techniques to maximize retention across multiple waves of data collection. We will discuss strategies for clear communication, such as crafting effective reminder messages, and how to structure incentives for sustained engagement.

By the end of this chapter, you will be equipped with the practical knowledge needed to set up and manage online data collection efficiently. Whether you are embarking on your first online study or seeking to refine your approach to more complex projects, the strategies outlined here will help you conduct your research with greater confidence and success.

Chapter 14 Outline: Best Practices for Setting up and Launching Online Studies

Module 14.1: Data Collection Best Practices



Reading

Review the best practices for online data collection and learn how to use them for complicated studies

- Learn how to successfully set up and launch an online study
- Read about best practices for describing studies and sampling participants
- Learn how to manage complex projects, like longitudinal surveys

Preview

Setting up and Launching an Online Study

Module 14.1

Walk through the steps to set up and launch an online project

Setting up an online study takes just a few minutes, but conducting quality research requires careful planning. This planning often involves knowing how to describe projects to participants, how to set appropriate compensation, target specific samples, and control sample composition. For longitudinal or more complex studies, it is often necessary to combine all these aspects to get the most out of participants and maintain data quality.

In the sections that follow, we review strategies for successfully collecting online data.

Best Practices for Setting Up an Online Project

The basic process of setting up an online study requires you to connect your survey to a participant recruitment site, describe the project, select eligibility criteria, launch the study, and manage incoming data. While each site varies somewhat in its user interface and capabilities, the basic principles of gathering data from paid sources of participant recruitment are largely the same. Here, we describe the best practices of setting up an online study with the CloudResearch Connect platform.

The typical setup process on Connect starts with a survey link from a platform like Qualtrics, SurveyMonkey, or Engage. Once the survey link is in place, the next steps are configuring who can participate and how they will be paid, and then monitoring responses as they come in. Throughout this process, you will make decisions that affect recruitment speed, data quality, and the participant experience.

If you have already created a Connect account—perhaps to participate in studies in Chapter 2 or to collect data in any of the previous chapters—you can log into the researcher portal using your existing credentials. If you are new to Connect, visit <https://bit.ly/3DnsQHI>. Click “Create an Account” and follow the prompts.

After logging in, you will see the dashboard with all your current projects (Figure 14.1). Simply click the blue “Create Project” button in the top right corner to begin setting up a new study.

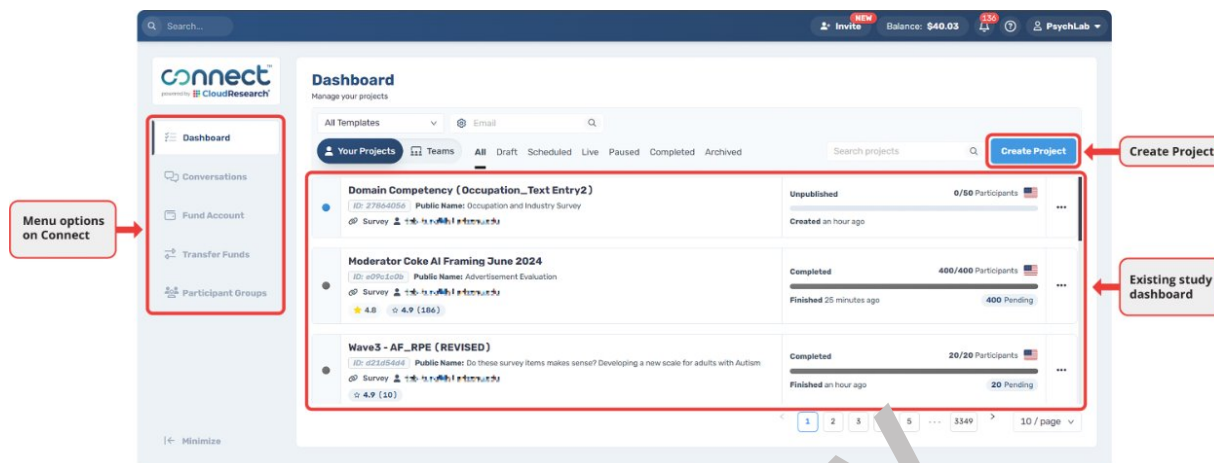


Figure 14.1. Connect dashboard showing navigation tabs on left and study listings in center.

How to Name a Study

The first step of any project is to give it a **project name**. Connect provides two places to name your project. The first one is a public name that participants will see on their dashboard when browsing studies (Figure 14.2). The second one is an internal name.

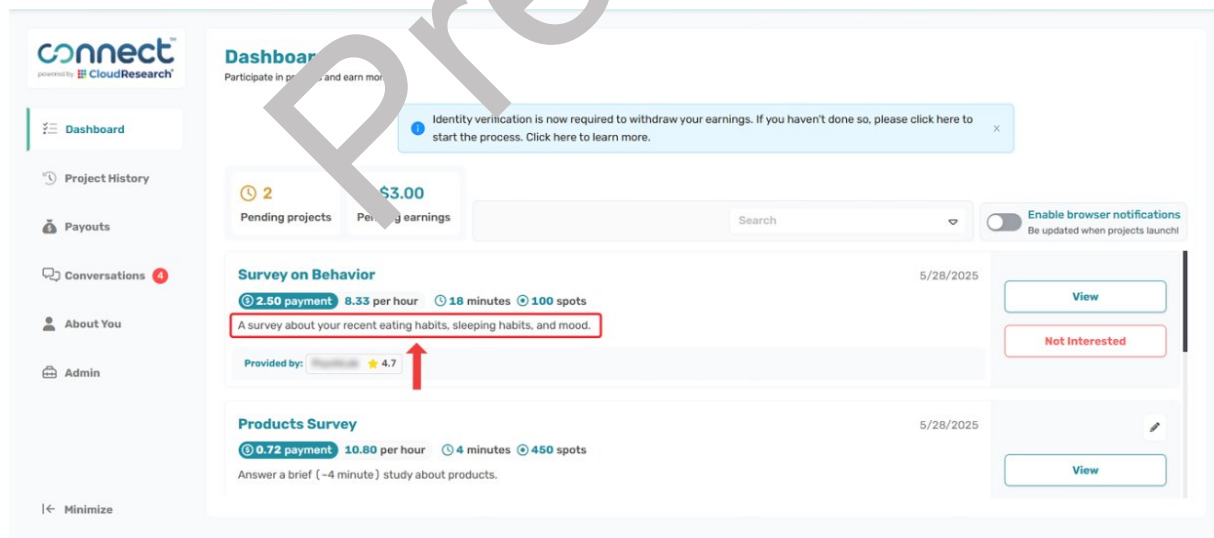


Figure 14.2. The text entered in the “project description” box appears on the participant dashboard as they browse projects.

💡 It is generally best to use generic wording like “Survey about attitudes” or “Research study in psychology” to avoid selection bias. For example, if your study is about people’s attitudes toward gun ownership, you should avoid titles like “A study about attitudes toward guns.” People with strong attitudes toward gun policy, on both sides of the debate, might be more (or less) drawn toward participating, creating the potential for selection bias.

The one exception to this rule is if the study has special requirements. These requirements should then be included in the study’s title and description, as appears in Figure 14.3. Notice how the study in Figure 14.3 has a very generic name (Academic Study), but before the name the researchers listed several requirements such as access to ChatGPT and Google Chrome. Failing to include this information up front will confuse participants and create significant dropout, as people will accept the study and only then realize they don’t meet the study requirements.

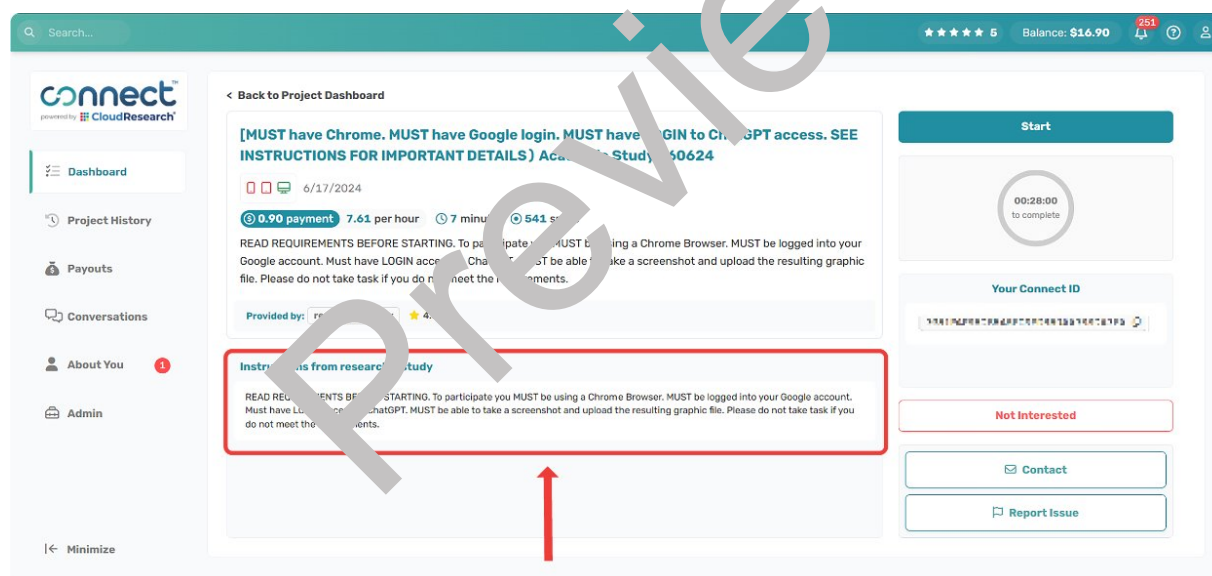


Figure 14.3. After choosing to view a project, participants are presented with special instructions from the researcher.

How and Where to Describe Your Project

After naming your project, Connect provides two places to describe it. The first place is a brief *project description box*. Here, you can briefly tell participants what your study is about (e.g.,

answer some questions about personality). This information will appear on the dashboard as they browse studies.

The second place to describe your project is in the instructions to participants box. This section offers expanded information and appears when participants click "View" before accepting the project (Figure 14.3).

💡 For simple surveys, you may provide minimal instructions (or perhaps no instructions).

However, when your project requires participants to download software, participate in video calls, view sensitive content, complete longitudinal follow-up surveys, or do anything more than answer survey questions, it is best to provide clear details within the instructions box (in addition to a brief mention of this in the title). Ideally, this information will be presented in a way that is easy for participants to immediately grasp, such as with bullet points or a numbered list.

The first thing to note about study descriptions is that participants do not read them very carefully. This is true for both the informed consent page for a survey and for the description page on Connect. Therefore, it is best to include only the most critical information in the study description field and to present it in bullet point form.

For example, if the study is very long (over 45 minutes), you could write: "This study requires at least 45 minutes to complete. Please only accept if you have 45 minutes right now." This discourages rushing and reduces dropouts.

For longitudinal studies, clearly stating that follow-up sessions are required—along with information about how long they will last and how much participants can earn—allows people to self-select into the study based on their willingness to return, improving retention rates (e.g., Hall et al., 2020). For other kinds of studies, a clear project description can be equally effective.

How to Anonymously Follow up with Participants

When you conduct an online study, the data is typically gathered through a survey platform like Qualtrics, SurveyMonkey, or Engage. The dataset, on its own, contains no personally identifiable information about the people who participated. However, there are many common research

scenarios where you, as the researcher, will need to link specific survey responses back to individual participants without compromising their anonymity.

Longitudinal Research: Longitudinal studies require tracking the same participants across multiple waves of data collection. This is impossible without a reliable way to link each participant's data from one wave to the next. Specifically, a study with two waves of data collection will have two datasets, and the rows of those datasets (corresponding to specific participants) need to be matched up. The only way to do that is with participant IDs.

Awarding Bonuses: You might design a study where participants receive a bonus payment for high performance. To deliver this bonus, you need to know which row in the dataset corresponds to the specific participant you want to bonus.

Managing Data Quality: If some participants fail attention checks or provide data that indicates fraud (see Chapters 10-11), you might want to prevent these people from participating in your future studies or to not pay them. This requires knowing the participant IDs.

Selective Recruitment for Future Studies: Sometimes a subset of your participants might meet specific criteria that make them ideal candidates for a follow-up study. To invite only these people, you need to be able to identify them.

Follow-up Questions or Clarifications: In some cases, you might need to follow up with a specific participant to clarify a response or to address an issue they reported, and a unique identifier makes this possible while still maintaining anonymity.

In all these instances, the ability to link study data to a unique participant identifier is required. CloudResearch Connect is designed to facilitate this with **anonymous participant IDs**. These IDs are unique alphanumeric strings assigned to each participant by the platform. They allow you to manage participation, track participants across studies, and distribute payments without needing to collect personal information like names or email addresses within your survey data, thus protecting participant anonymity.

The primary method for linking these anonymous IDs to your survey data is by embedding the participant's Connect ID into the data file. There are two ways to do this. The simplest way is to ask participants to paste their Connect ID into the survey. This will create a column of participant IDs that can be copied and pasted to Connect for longitudinal follow-up, bonusing and other forms of participant management. However, this method is prone to errors. Some respondents will not copy their ID accurately or not comply with the request for some other reason.

The less error-prone way to add participant IDs to the datafile is to configure the study on Connect to automatically pass each participant's unique Connect ID to the survey when they click the study link. Connect provides platform-specific instructions for using what is called "embedded data" to automatically record this information as a variable in your dataset.

After providing a link to your survey, you can configure the study to capture each participant's Connect ID (Figure 14.4). The system provides platform-specific instructions for using embedded data to automatically record this information.

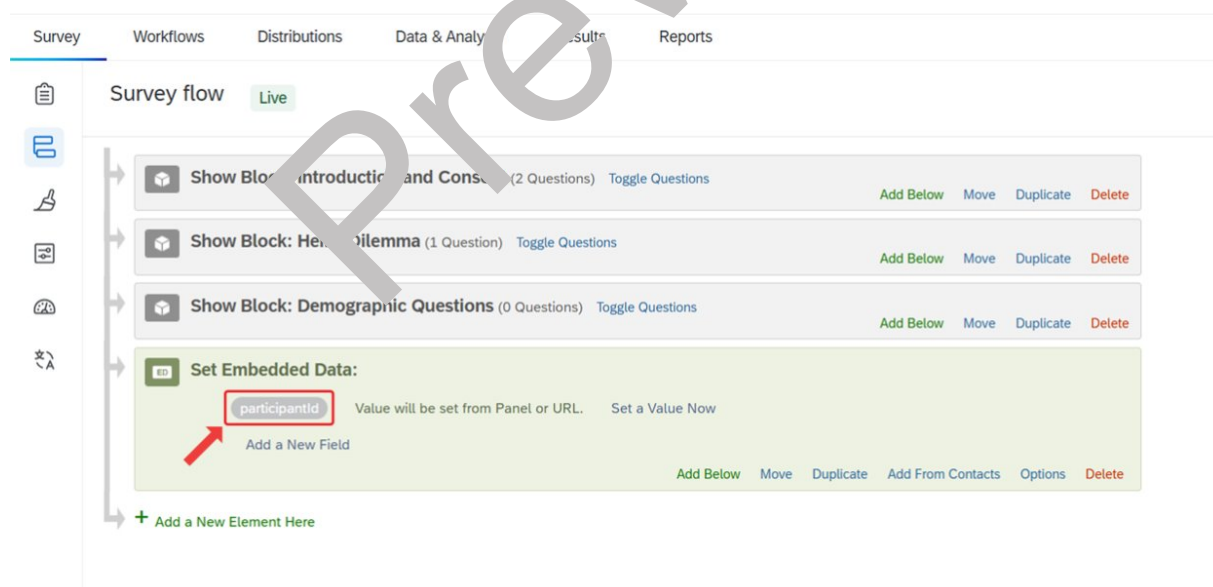


Figure 14.4. Embedded data allows you to capture people's participant IDs without asking for them in the survey.

How to Figure out Fair Payments

Fair payment starts with understanding platform norms. On Connect, most researchers pay around \$10 per hour, and participants consider \$11 per hour fair payment (Moss, 2024).

Although the average payment is around \$10 per hour, there is a range to what researchers can pay. Often, the amount that is appropriate depends on the project and what participants are asked to do.

The minimum pay researchers can offer on Connect is \$7.50 per hour and the recommended payment is around \$8.50 per hour for basic surveys and simple tasks. This means for a typical 15-minute study, participants earn about \$1.85 and \$2.15.

As projects become more complicated, however, payment should increase. When participants are asked to devote extended attention or effort, engage in creative thinking, solve complex problems, provide personal information, use special skills or knowledge, return for multiple rounds of data collection, or engage in something like a diary diary study, they should be paid more than the minimum rate. In a mock jury trial, for example, participants may need to devote an hour to closely reading both sides of a lawsuit and then render a verdict. For these projects, participants often earn \$15-20 per hour or more.

Finally, it is almost always a good idea to conduct a pilot study to accurately measure the actual completion time of a study.

How to Find the Participants you Need

Many research studies aim to draw samples from the general population. However, other studies have more specific recruitment goals. These might include people with certain political leanings, like Democrats or Republicans, those who have specific conditions like symptoms of depression or past suicidal ideation, or people within defined age brackets or socioeconomic status groups. There are several methods for selectively recruiting such specific groups of respondents.

The first and often most straightforward method involves using pre-existing qualifications offered by the platform. For example, CloudResearch Connect maintains profiles on its participants based on hundreds of questions they answer when they join the platform and on an ongoing basis. These qualifications cover a wide range of demographic, behavioral, and

attitudinal variables. If your research goal is to sample only Republican participants, for example, you can select this specific political affiliation qualification when setting up your study. Participants who have previously identified as Republican would then be eligible to see and participate in your study, while others would not.

Figure 14.5. The Connect interface allows researchers to select specific characteristics for recruiting participants.

It sometimes happens that researchers need to recruit a group for which a standard qualification does not yet exist on the platform. In such cases, there are generally two approaches. One option is to request the qualification and submit a form detailing the criteria they wish to target (see Figure 14.5). Connect will then add this as a new qualification, often making it available within a few days for use.

A second approach is to use what are referred to as "branches." Branches allow you to create your own custom screening questions that all potential participants answer when they start the study. For those who qualify by answering in a specific way (e.g., indicating that they "Eat avocados at least twice a week"), they become eligible and continue to the main part of the study.

For those who do not meet your criteria participation ends after the screening questions. Importantly, participants who are ineligible are still compensated for the brief time they spend answering the initial questions at the hourly rate advertised for the study. This provides a way to quickly qualify respondents based on specific criteria, when those criteria are not available as standard platform qualifications.

Creating Quotas.

Researchers often need more fine-grained control over their sample composition than simply including or excluding a specific category of people. For example, while you might target participants from a specific political party, you may also want to recruit several political party groups and have them appear in specific proportions within your sample. This is where quotas become highly useful. Quotas allow you to define specific "bins" of subgroups based on one or more characteristics and then recruit participants until a predetermined percentage or number for each bin is met (see Figure 14.6).

The screenshot shows a 'Demographic Targeting' interface with three steps: 1. Demographics, 2. Quotas, and 3. Summary. Under 'Political Party', there are three quotas, each with a count of 100 and a percentage of 33.3%. The categories are Democrat, Republican, and Independent. A red box highlights the 'Republican' category, and a red arrow points to the '100' count field. Another red arrow points to the '33.3%' percentage field. At the bottom, it shows 'Total: 300' and '100%'. There are 'Previous' and 'Next' buttons.

Quota #	Category	Count	Percentage
Quota #1	Democrat	100	33.3%
Quota #2	Republican	100	33.3%
Quota #3	Independent	100	33.3%
Total		300	100%

Figure 14.6. Each bin or bucket within a quota can be adjusted to the desired percentage of the sample. In this case, the sample will be 33.3% Republican, 33.3% Democratic and 33.3% Independent.

For example, imagine you are conducting a study in which you want to recruit an equal number of Republicans, Democrats, and Independents. Setting quotas for these bins will make sure that that you do not end up with a sample heavily skewed towards one or another political party.

On Connect, once you have selected a demographic targeting variable (like age, political affiliation, or income level), you can simply toggle an option to set quotas for that variable.

After enabling quotas for a chosen variable, you then define your "bins." For each bin, you can specify the desired percentage of the total sample that should fall into that category (e.g., 33.3% for the Republican bin). The platform will then manage recruitment to fill these quotas. For example, Connect will stop recruiting Republicans once that target is reached, even if other quotas are still open.

How to Increase Representativeness.

Quotas are particularly valuable when the goal is to increase sample representativeness. Connect, offers preset templates matched to U.S. Census for age, gender, race, and ethnicity. By selecting "Apply Census Matched Template," these quotas are automatically added to the project, (Figure 14.7). You can then add additional criteria such as geographic region, education and many others if needed.

The screenshot displays the 'Quotas' section of the Connect platform interface. At the top, there are buttons for 'Apply Census Matched Template', 'Request a Demographic', and a '5' icon. Below these are three tabs: 'Demographics' (selected), 'Quotas', and 'Summary'. The 'Quotas' tab is active, showing a section for 'Age' with the question 'What is your date of birth?'. A 'Quota Targeting' toggle is turned on. There are four quota bins, each with a slider for 'Min' and 'Max' age, a person icon, a '66' value, a slider, and a percentage target. The total sample size is 'Total: 300' and '100%' is reached. A '+ Add Quota' button is at the bottom left.

Quota #	Min	Max	Person Icon	Value	Target %
Quota #1	18	29	66	22	%
Quota #2	30	44	78	26	%
Quota #3	45	59	78	26	%
Quota #4	60	99	78	26	%

Total: 300 100%

Figure 14.7. The “Census Matched Template” adds quotas matched to the latest US Census figures for age, race, ethnicity, and gender.

When is the Best Time to Launch a Study

Once you have named and described the study, added a study link, and set the recruitment criteria, you are ready to launch the study. But when is the best time?

A significant advantage of online research is its lack of time constraints. Unlike laboratory studies, researchers can launch projects at any hour. Likewise, participants can complete studies at their convenience. Most studies on Connect and Mechanical Turk peak between 8:00 am and 6:00 pm (see Litman and Robinson, 2020), but participants take studies twenty-four hours a day, seven days a week. However, the participants who are online late at night may be significantly different than those who take studies during the day (e.g., Arechar et al., 2017; Casey et al., 2017; Fordsham et al., 2019).

Participants active late at night report lower levels of conscientiousness and higher levels of anxiety, depression, procrastination, internet compulsions, disruptive sleep behaviors, disordered eating, and neuroticism compared to those who are active during typical daytime hours (Fordsham et al., 2019).

If there are theoretical reasons to believe that the variables mentioned above are correlated with your research topic, and may thus potentially introduce some level of bias, it is best to carefully control your launch time.

Once a study is ready to launch, there is an option to schedule a specific launch time using the calendar wizard to select your desired date and time. Your project will then automatically go live at the specified time and will appear on your Dashboard as “Scheduled.” Even if you are working late at night to create the study, it is best to use the scheduler to set the study to launch in the morning or afternoon.

Once launched, your project becomes visible to eligible participants based on your targeting criteria, allowing data collection to begin.

How and When to Communicate with Participants

The Conversations center in Connect operates like an email system while protecting participants' anonymity (Figure 14.8). Researchers can message individual participants or groups by entering Connect IDs into the recipient's box. This system of communication helps when addressing technical issues, explaining a rejection, awarding a bonus, or sending reminders about follow-up waves in longitudinal studies. Participants can also contact researchers to ask questions or explain issues when completing a study.

It may go without saying, but communications should always be polite and professional. As the anecdote at the start of the next Chapter will demonstrate, polite communication shows respect for participants' time while injecting some humanity into a largely impersonal transaction.

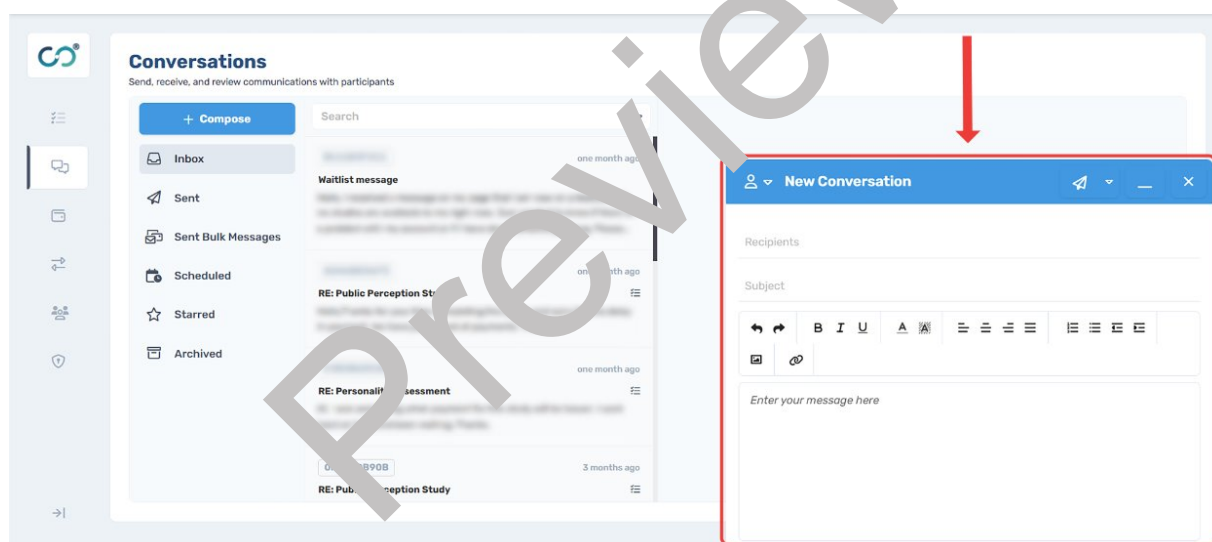


Figure 14.8. The conversations center allows researchers and participants to communicate with each other while maintaining participant's anonymity.

Best Practices for After a Project is Launched

When a project is launched, Connect creates a list of qualified participants and sends them notifications. These notifications speed up data collection when compared to simply publishing the study without alerts (Litman & Robinson, 2020). Additionally, email notifications pull in less active participants who might otherwise miss the study, as they are less likely to be monitoring

their dashboard. Because less active participants receive higher priority in the notification system, studies typically fill with more casual participants than would otherwise be the case (see Robinson et al., 2019).

Regardless of what happens behind-the-scenes, once you launch a project, data will begin streaming in. Therefore, it is important to know what to do when a project is live.

What To Do While a Project is Live

While a project is live, the dashboard provides real-time monitoring. You can see how many people are actively working on the project, the number of completed submissions, and the bounce rate (Figure 14.9).



Figure 14.9. When a study is live, the project dashboard displays information that is useful for monitoring study progress.

The bounce rate is particularly valuable. It refers to the percentage of people who start the study but then return it. The typical bounce rate on Connect is close to zero or in the low single digits. A bounce rate exceeding 20% typically signals an issue that requires attention. If the bounce rate surpasses 50%, Connect automatically pauses the project to prevent a negative participant experience.

Several factors can cause a high bounce rate, such as a project description that fails to set accurate expectations, a technical problem such as a broken link, or a survey element that participants find confusing or objectionable.

During the study, participants can provide feedback that might indicate the source of the issue. Connect provides two feedback channels. The first is post-completion reviews. After finishing a study, participants can rate their experience, including the accuracy of the time estimate, the

fairness of the compensation, and the researcher's conduct. They can also provide open-ended comments that may alert you to any problems.

Second, Connect offers technical error reports. Through this channel, participants can report specific problems with question phrasing, answer options, programming mistakes, downloads, broken links, or media playback issues (Figure 14.10).

Participant ID	Category	Details	Reported At
A2937D9968	Policy Violation	This survey made me sign a petition.	6/17/2024 - 4:34 PM
7E8CEF3E1A	Technical Issue	Your links are broken and then sends me all the way back to the beginning of the survey. This is too long to go to the end only to get sent back to the beginning with no payment.	6/17/2024 - 12:40 PM
9380754808	Technical Issue	couldn't go back to the survey after checking out a website for AI	6/17/2024 - 10:55 AM
4E860088D5	Broken Content	I was close to the end of the study when I clicked on the suggested observation for AI developments and I wasn't able to return to the study.	6/17/2024 - 10:50 AM
96660CDCD4	Technical Issue	After the petition signage I looked at website about AI and it would not let me out of this to finish and get my reward cord, not sure if that was the	6/17/2024 - 10:48 AM

Figure 14.10. Reports of broken content or other technical errors from participants can be invaluable, sometimes preventing scarce resources from being wasted on a project that contains errors.

Monitoring these feedback channels, especially during the first few minutes after launch, can help identify and address problems before they compromise your entire dataset. If participants report issues that need correction, you should pause the project on your dashboard. Pausing will prevent new participants from accepting the study while allowing those already engaged to complete it. Depending on the issue, you may need to make changes to your survey or Connect settings before resuming data collection.

If there are no problems when you launch your project, all you need to do is sit back and watch the data roll in.

General Best Practices

Pilot Testing

Pilot testing involves opening a study to a small number of participants as a trial run before collecting a complete dataset. Although many research projects operate on tight timelines, pilot testing is worthwhile because it allows you to verify that data is being properly recorded, catch programming errors or problems with study materials, check that links and survey logic are functioning correctly, accurately estimate completion time, and determine appropriate compensation.

Connect makes pilot testing simple. When launching a project, you will be asked whether you want to conduct a pilot or launch the full study. A pilot launch with even a few participants can reveal issues you might have overlooked. Launching a pilot with ten or twenty participants is a great start. After the pilot is complete, you can smoothly transition to the full launch.

Fixing Study Errors

What should you do if you find a mistake during the pilot phase or while the study is live? First, you should pause the study by clicking on the “...” menu and selecting “Pause.” Pausing a study prevents new participants from starting it, but anyone currently active will be given the chance to complete their session.

Once the study is paused, you can correct the error. This usually means changing something in the Qualtrics survey or editing the payment, study length, or task description on Connect.

If your changes take place in Qualtrics, you can simply make the changes, save them within your survey platform, and then resume data collection on Connect. When analyzing the data you will want to make sure you separate people who experienced the error from those who did not, but otherwise your data collection can continue within the same project. If you need to change parts of the study on Connect, however, you will need to close the study and then clone the project. Cloning a project copies all your study details into a new project where you can edit them. In

addition, cloning ensures that each participant from the original project is ineligible for the second project so you can avoid duplicate participation.

After editing a project, you should double-check the changes are complete and then pilot the new project before moving to a full launch.

Conducting Successful Longitudinal Studies

Online research can be used to conduct a variety of complex projects, including video interviews, interactive experiments with multiple respondents interacting with each other, at-home product testing, and many others. Of all the complex types of designs, longitudinal research is perhaps the most common.

Longitudinal studies track the same participants over multiple sessions. We have encountered longitudinal designs multiple times throughout this book. In Chapter 11, we analyzed a longitudinal dataset with five hundred participants who completed anxiety, depression, trauma, and other measures on two occasions over one year apart. As we discussed in-depth in that chapter, longitudinal studies allow researchers to build a stronger case for causal inference by establishing temporal precedence. Longitudinal research also allows researchers to examine developmental changes, the trajectory of attitudes or behaviors over time, and the long-term impact of events or interventions. Many research questions in behavioral sciences are inherently longitudinal. For example, assessing the test-retest reliability of a measure requires having the same respondents come back to participate on at least two separate occasions.

For these reasons, longitudinal studies are extremely common online. This is because online platforms make it much easier to recruit, track, and retain participants across multiple sessions.

However, the success of any longitudinal project hinges critically on participant retention.

Losing participants between waves of data collection can reduce statistical power, introduce bias if those who drop out differ systematically from those who remain, and ultimately undermine the study's validity.

Chandler et al., (2021) examined 1,200 online longitudinal studies spanning numerous behavioral disciplines and including over 36,000 participants. They provided recommendations for best practices when launching longitudinal studies online. By applying these and other

recommendations for best practices (see Hall et al., 2020) researchers can significantly improve retention and the overall quality of their longitudinal data.

The Immense Potential of Crowdsourcing Platforms for Longitudinal Studies

Generally speaking, longitudinal retention in online studies is very high compared to more traditional sources (see Hall et al., 2020). The average retention on MTurk across 1,200 separate studies was close to 70%, with much higher retention over short intervals. Similar and even higher retention rates can be expected on Connect and other crowdsourcing platforms. However, retention across longitudinal sessions is significantly lower on market research panels. Typical retention across one week is only around 30% and falls even more with a longer gap between intervals.

For most longitudinal studies, crowdsourcing platforms are preferable to other alternatives, such as market research panels, volunteer panels, and university subject pools. Indeed, beyond relatively simple two wave longitudinal studies, crowdsourcing platforms allow for longitudinal designs that are not feasible anywhere else. For example, in a study conducted by IARPA, two thousand CloudResearch participants engaged in 2-hour political forecasting tasks every week for over one year (Moss, 2022). Eighty-five percent of these respondents completed close to 90% of the sessions over that period. As discussed in Chapter 9, OpenAI ran a study in which 1,000 respondents interacted with ChatGPT data for a period of thirty days. And in one study, the authors of this book conducted a ten-year follow-up in which 30% of people came back ten years after the initial session. As all these studies demonstrate, the possibilities of using online platforms for longitudinal recruitment are immense and likely provide opportunities for longitudinal follow up beyond any other participant recruitment source.

Study Title and Description

The groundwork for good retention in a longitudinal study begins with how the study is initially presented. As emphasized in the previous section on setting up studies (see Figure 14.3), it is important to clearly state the longitudinal nature of the project in both the study title and the detailed description. Participants should be informed upfront about the expected number of waves, the approximate timing or frequency of each follow-up, the estimated time commitment per wave, and any specific requirements for ongoing participation. This will allow people to self-

select into the study based on their willingness and ability to commit to the entire project, which can significantly improve long-term retention rates.

Compensation and Incentives

Compensation plays a significant role in improving retention rates. Offering people above-average payment increases retention in online longitudinal studies (Hall et al., 2020).

Additionally, increasing the incentive for each successive wave to reward continued participation and offering a completion bonus for finishing all waves of the study are recommended practices.

In one study, Hall et al., recruited 600 geographically diverse participants for seven sessions that spanned one year. The payment was increased incrementally across sessions from 50 cents to \$1, \$2, \$3, \$4, \$5, \$6, to \$7 with an additional \$5 bonus for completing the last session. Their retention rate was 73% for the second session, but after that, retention remained very high (around 95%) throughout the rest of the study.

Communication and Reminders

To successfully conduct longitudinal research, it is important to communicate with participants.

When researchers remind participants about upcoming study sessions and encourage participation, they often see less attrition than researchers who do not send these messages (Chandler et al., 2021). Often, a polite, timely reminder about participating in a study is enough to increase participation and minimize attrition. These reminders also help participants stay engaged and aware of upcoming sessions.

Reputation Qualification

When maximizing retention is of utmost importance, participant qualifications can be helpful. As discussed earlier, platforms like Connect allow researchers to target participants based on various characteristics. Selecting participants who have a track record of active participation on the platform can increase the likelihood that they will return for future waves. In fact, highly active participants are more than 60% less likely to attrit compared to less active participants (Hall et al., 2020).

Tracking IDs

Effective tracking of participant IDs is also important for longitudinal research. As described above, embedding anonymous participant IDs into the dataset allows you to match data from the

same participant across multiple waves. This is essential not only for data analysis but also for managing communications, sending targeted reminders to those who have not yet completed a current wave, and for accurately distributing any wave-specific payments or completion bonuses.

Facilitating Large-Scale and Intensive-Longitudinal Designs

Behavioral scientists are increasingly interested in large-scale longitudinal research. This may mean tracking the same participants over an extended period or asking people to complete intensive-longitudinal research, such as experience sampling and daily diary studies (Bolger and Laurenceau, 2013).

Managing these complex studies presents significant logistical hurdles. Imagine a daily diary study requiring three data collection points each day for a month. This could mean setting up and managing 81 separate online studies. This includes figuring out who is eligible for each wave (depending on prior participation), sending reminders, tracking responses for each specific wave, and implementing potentially complex incentive structure based on performance or completion of multiple stages.

The Connect platform offers a feature called “waves” that is specifically designed to automate the setup and management of such studies. Researchers can use a setup wizard to configure the entire study timeline at once. This includes specifying the number and frequency of all data collection waves, defining when each wave should launch, determining which participants are eligible for each stage, and setting the compensation structure, including potential bonuses for completing a certain number of waves. By using these tools, researchers can conduct more complex studies investigating how social and psychological phenomena change and unfold over time without the hassle of manually administering the study.

Summary

This chapter reviewed practical strategies for effectively setting up and launching online research studies. We examined best practices for naming and describing a project to minimize selection bias, the importance of clearly communicating any special study requirements, and methods for anonymously tracking participant IDs. We also covered how to determine fair compensation based on platform norms and study complexity, and how to use platform tools for targeting specific participant groups and applying demographic quotas, such as those matched to the U.S.

Census, to enhance sample representativeness. We also discussed considerations for the optimal timing of study launches and the importance of real-time monitoring of data collection, including how to interpret participant feedback and address technical issues promptly.

We placed a strong emphasis on conducting successful longitudinal studies, focusing on strategies for increasing participant retention. Key strategies discussed included setting clear expectations about the study's duration from the outset, staggering the compensation and incentives to encourage retention, selecting participants who are more active in the platform, and the role of consistent communication, such as sending reminders for follow-up waves. We also explored how platform features, like Connect "Waves," are streamlining the management of large-scale and intensive-longitudinal designs.

Throughout this chapter, a central theme has been that successful online research requires thoughtful and participant-centered practices. The practices apply regardless of the specific online platform used or the precise nature of the research question being asked. They are also the focus of the next chapter on research ethics and the participant experience.

Chapter 15: Ethics and the Participant Experience

Participants sometimes contact researchers or the platform they are using to report a problem with a study. In 2023, the CloudResearch team received one of these messages. The person was upset because they had been rejected from a study. The rejection meant the participant would not be paid for their time. They felt the rejection was unwarranted, so they contacted us to voice their frustration.

Our team quickly replied. We informed the participant that CloudResearch would contact the researcher, ask them to double check their records, and explain why the rejection was warranted. Within minutes of receiving our reply, the participant wrote back.

“I feel better now, no matter what the outcome is! Thank you so much for listening to me and you have a great rest of the day...and THANK YOU SO MUCH for the nice reply.”

Notice, nothing changed. Our email didn't overturn the rejection, nor did it promise payment. CloudResearch didn't take any sides, and the participant didn't instantly receive an explanation for the rejection. All those things came later. So, why did the participant feel better? We think it's because someone listened.

Stories like this are a reminder that behind every online data point there is a person. The people who take part in online studies invest their time and emotions into the work. When researchers treat participants with care, everyone benefits—scientists obtain quality data and participants feel valued for their contributions. But when the relationship between researcher and participant breaks down, participants feel dismissed and disrespected. In the worst cases, they lose their time, their money, or faith in the research process.

This chapter examines the ethical considerations relevant for conducting online research. Module 15.1 begins by examining the historical context that shaped current ethical guidelines, focusing on the Belmont Report and its three core principles: Respect for Persons, Beneficence, and Justice. It then illustrates how these principles translate into practical applications such as informed consent, risk/benefit analysis, and fair participant selection, specifically in online contexts. Finally, the module shifts to the practical application of these ethical principles in the

day-to-day conduct of online research. We will address practical questions such as how to set fair compensation, the responsible handling of participant rejections, ensuring participant privacy and data confidentiality, and the importance of clear communication.

After discussing the basics of research ethics, Module 15.2 provides a guide for preparing a research ethics proposal. This module contains templated answers to common IRB questions and is intended to help anyone filling out an IRB application for online studies. Ultimately, this chapter aims to help you conduct online research that is not only scientifically rigorous but also respectful and considerate of the participants who make research possible.

Preview

Chapter 15 Outline: Ethics and the Participant Experience

Module 15.1: The Basis of Ethical Research



Reading

Explore the background of ethics in human subjects research

- Learn where the foundations of ethical research come from
- Explore how the principles of ethical research are applied in online studies

Module 15.2: Conducting Ethical Research: Writing an IRB Application



Reading

Examine a sample IRB application and learn how to complete your own research proposal

- Work through the questions in a typical IRB proposal
- Learn how to answer common concerns IRB's have about research

The Basis of Ethical Research

Module 15.1

Explore the background of ethics in human subjects research

All colleges, universities, and research hospitals in the United States that accept money from the federal government are required by law to have an institutional review board (IRB) that ensures the ethical conduct of research involving human participants.

An IRB, or ethics board, is a committee of people who review research proposals to make sure studies meet ethical guidelines. Before a researcher can conduct a study, they must describe their methods, strategy for recruiting participants, any potential risks to participants, and how risks will be mitigated to the IRB. The IRB reviews the proposal and either approves the study, requests modifications, or rejects the proposal if the risks to participants are unacceptable.

The criteria IRBs use to evaluate research stem from the Belmont Report, a foundational document in research ethics created in the 1970s by a diverse commission of scientists, ethicists, and citizens. The Belmont Report was created in response to significant ethical failures in research such as the Tuskegee Syphilis Study, in which the U.S. Public Health Service observed Black men with syphilis for several decades, deliberately withholding effective treatments like penicillin. This unethical conduct led to preventable suffering, the continued spread of the disease, and loss of life. Such events underscored the need for formal ethical guidelines to protect research participants.

The Belmont Report established three core principles for human subjects research: **respect for persons**, **beneficence**, and **justice**. These three principles are translated into applied research practices, as illustrated in Figure 15.1.

For example, the ethical principle of respect for persons means that people should be treated as intelligent beings who can decide if they want to participate in a research project or not. A practical application of this principle is **informed consent**. Informed consent is the process of telling participants what they will be asked to do in a study, what the risks and benefits of participation are, and asking if they agree to participate.

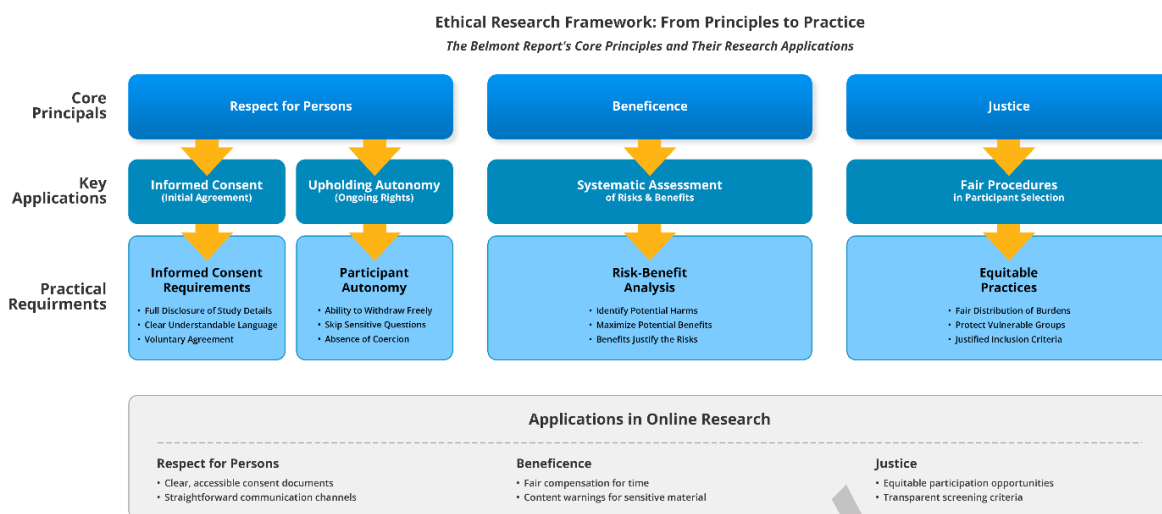


Figure 15.1. A depiction of how the core principles of the Belmont Report get applied in specific parts of the research process.

A second principle from the Belmont Report is *beneficence*. Beneficence involves protecting participants' well-being by weighing the risks of a study against its potential benefits. This is one of the main things IRBs consider when evaluating research. In online research, this principle includes protecting participants' data and online privacy.

The third principle from the Belmont Report concerns *justice*, which calls for fairness in the selection of research participants and in the distribution of research benefits.

While the Belmont Report principles are foundational across all research, it is important to recognize that most contemporary online studies present *minimal risk* to participants. Thus, from the perspective of the principle of beneficence, the risks (which are minimal) are typically justified by the benefits in a risk/benefit analysis. For this reason, the overwhelming majority of online studies qualify for *exempt review*—the lowest level of scrutiny an IRB directs toward research projects. At the same time, some aspects of online research can be distressing to participants.

This chapter will explore how the ethical principles of research are applied in online studies. Our discussion will begin with what is currently understood about the risk of harm in online studies, including those that involve research on sensitive topics.

Risk of Harm in Online Research Studies

The principle of beneficence is a cornerstone of the Belmont Report. It obliges researchers to secure the well-being of participants by maximizing potential benefits and minimizing potential harms. This requires a careful assessment of the risks involved in any research study. The first step in this assessment is to identify potential sources of harm.

What might harm participants in the online research environment, and what level of risk do those harms pose to participants? Historically, most online research has been considered to involve minimal risk (e.g., Kraut et al., 2004). Most online studies consist of questionnaires, vignettes, or tasks measuring cognitive processes like memory, attention, and reaction time. Engaging with such stimuli generally does not provoke more stress than what someone might encounter in everyday life, which is often the benchmark for evaluating harm in behavioral research.

Large studies conducted on platforms like Mechanical Turk have asked participants whether participating in online studies is more stressful than everyday life (Litman & Moss., 2020; Moss et al., 2023). One study with over 10,000 MTurk participants who had completed hundreds of thousands of studies found that just 4% of people indicated that participating in online studies was more distressing than daily experiences. And even among this 4% of participants, most (75%) said that the benefits of participating in research outweighed the risks. These findings suggest that online research generally poses minimal risk for participants. Furthermore, because these participants had collectively completed hundreds of thousands of studies conducted by thousands of different researchers, it further suggests that online research studies are often benign.

At the same time, it is important to understand why a small percentage of participants reported that research was sometimes distressing. By understanding these participants and their experiences, researchers may be able to improve the experience of all participants.

To understand what people find distressing about research, the same study followed up with the ~4% of participants who reported that online studies were more distressing than daily life. It asked them to describe the sources of their distress. The answers fell into five categories: 1) Unclear instructions, 2) Unfair rejections, 3) Uncommunicative researchers 4) Competition for

studies, and 5) Upsetting or inappropriate study materials. Table 15.1 lists these categories along with relevant quotes from participants.

In the next several sections, we discuss how researchers can structure studies to minimize each source of participant distress, starting with research on sensitive topics.

Category	Participant Quotes
Fear of Rejections	<p>"Avoiding rejections, getting rejections regardless of honest/good quality work."</p> <p>"Finding requesters that don't reject for a simple mistake."</p> <p>"Knowing you didn't miss an attention check, but getting a rejection anyway."</p>
Unfair Treatment	<p>"Most researchers are nice people, but some will just not respond or threaten people with blocks or rejection. Some researchers will lie about their pay rate."</p>
Sensitive Study Content	<p>"Occasionally I'm asked to watch and comment on a video that is upsetting. I find some of the videos upsetting for personal reasons, and I'm even more upset that they don't contain a content disclaimer before I try them."</p> <p>"Some times I have to look at and react to graphic pictures."</p>
Unclear Instructions	<p>"Unclear instructions."</p>

Intense Competition	"MTurk has hired so many workers that it takes as much time to find work as it does to actually do it. HRs literally disappear in milliseconds rather than hours."
	"Finding surveys to do is nearly impossible due to intense competition from others wanting to grab the same surveys . So in the end you end up with less to do unless you catch it first."
	"Good surveys being posted one at a time and disappearing instantly; not being able to make a living wage, underpaid work."

Table 15.1. Examples of challenges participants face in online research.

Research on Sensitive Topics Online

One area of concern within the risk/benefit assessment that IRBs conduct involves research on sensitive topics, such as trauma, sexual abuse, or other potentially upsetting experiences. A common assumption is that asking participants about such topics might increase the risk of harm or even retraumatize people, especially in an online setting where researchers cannot provide immediate support.

However, research studies, including studies conducted online, suggest that for most people participating in research that touches on sensitive or traumatic experiences is not significantly more distressing than everyday life (Newman & Kaloupek, 2004; Cromer & Newman, 2011; Newman & Kaloupek, 2009). Studies with survivors of trauma, for instance, show that people generally feel the benefits of the research outweigh discomfort, often expressing willingness to participate even if they had foreknowledge of the study's content (Griffin et al., 2003). Indeed, as reported in a literature review of this topic "the risks associated with asking about abuse are overstated and inconsistent with actual data" (Gleaves et al., 2007).

Nevertheless, when online participants report distress related to study content, it often involves exposure to graphic images such as violence, gore, sexually explicit content or painful personal experiences. This can be particularly upsetting to participants when such content is presented without adequate warning. A significant portion of participants who report distress from such content also indicate that clear warnings beforehand would mitigate their negative experience.

To conduct research on sensitive topics ethically and effectively in online environments, and thereby uphold the principle of beneficence, researchers should consider adopting the best

practices outlined below. Specifically, we recommend four actions that are aligned with the principles of ethical research.

Clearly communicate the nature of the study in and out of the consent form. Provide participants with a clear notice about potentially disturbing content before they begin the study. This allows participants to make informed decisions about their participation. The notice should be included in the consent document, but it should also be in the study description field on the recruitment platform. The reason is that most people don't read consent forms. For example, researchers in one study embedded phrase "some researchers wear yellow pants" within their consent form and most participants did not notice it was there (Douglas et al., 2020). This study occurred in-person in a laboratory, so the odds that people are reading online consent forms are even lower.

Allow participants to skip sensitive content. Allowing participants to skip sensitive questions is an easy and practical solution to mitigate against the possibility that some people will be upset at being forced to answer certain questions. Our experience shows that few people use this option, but the presence of an option to skip questions reduces people's stress without compromising data quality.

Conduct thorough pilot testing. Testing study materials and gathering feedback from participants allows researchers to identify and address potentially distressing issues before conducting a full study.

Provide access to support. If a study may upset participants, researchers can provide clear directions for support resources. While these resources are rarely needed, having these resources available fulfills the researcher's obligation to look out for participant's welfare.

Fair Payment and Rejecting Participants

Some of the most common sources of frustration for participants are fear of having studies rejected unfairly and low pay rates. For instance, participants often report that for a simple mistake, without clear reasons, or even when they believe they have provided quality work their

study will be rejected and the payment withheld. Participants also report that they perceive some researchers to be dishonest, most commonly regarding stated pay rates.

How Much Should a Study Pay? The question of fair compensation has generated more debate than any other aspect of online research ethics (see Moss et al., 2023). This debate centers on two issues: whether online research requires paying people minimum wage, and what constitutes fair payment when participation is voluntary.

A majority of people who participate in survey studies report that it is something to do in their leisure time (Moss et al., 2023). An analysis of tens of millions of tasks completed on Mechanical Turk over nearly a decade shows most researchers offered baseline compensation of 12-13 cents per minute, equivalent to \$7.25-\$7.80 per hour (Moss et al., 2023). More recent data from the Connect platform indicated higher rates, averaging \$10 per hour. Notably, participants themselves report that hourly wages near \$11 are fair compensation (Moss, 2024).

Based on this evidence and practical considerations, we recommend a baseline rate of payment of at least \$8.50 per hour for standard survey research. This rate acknowledges the budget constraints many researchers face while remaining within established norms. For context, a typical 10-minute study with 300 participants would cost approximately \$530 including platform fees.

This baseline rate should be adjusted upward for studies that are longer, more complex, or require special participant engagement. Higher compensation for demanding tasks not only reflects greater participant effort but also tends to improve data quality and participant retention, particularly in intensive longitudinal studies (see Litman et al., 2020).

Rejecting Participants. The ability to reject low-quality submissions presents a complex ethical challenge in online research. While it may appear on the one hand that rejecting participants violates the principle of respect for persons—traditionally understood to encompass voluntary participation and a participants' right to skip any question they are not comfortable answering—on the other hand, this view fails to account for the unique nature of online research environments.

Online research platforms rely on quality control mechanisms to maintain *the integrity of* research. When participants submit fraudulent data or fail to comply with the study's instructions, researchers need the ability to reject the submission. This serves three important functions: it preserves research resources, it maintains data quality, and it helps platforms identify participants who consistently provide poor quality data.

The need for quality control is especially apparent in complex studies. Consider a longitudinal study where participants receive premium compensation for sustained engagement, or specialized studies recruiting professionals at a higher pay rate. Without the ability to screen for quality through rejections, these expensive studies are impossible to conduct. To make sure that the power of rejection is exercised carefully, we recommend three practices.

First, establish clear criteria for quality before data collection begins. This prevents rationalization after the fact and ensures fair treatment of all participants. The discussions in Chapters 10 and 11 are particularly relevant for establishing such criteria.

Second, use rejections sparingly and only in clear cases of non-compliance, misrepresentation, or fraud. When participants appear to give genuine effort but the results are questionable, err on the side of approval. Then, you can use alternative methods such as flagging the participant or selectively excluding them from future studies to improve data quality.

Third, conduct thorough pilot testing. Pilot testing helps identify potential sources of confusion, allowing you to create clear instructions and an intuitive study design that naturally minimizes the need for rejections while improving overall data quality.

Protecting Privacy

Participants often report that requests for highly personal information are a source of negative experiences. However, online research platforms offer participants anonymity as they participate in studies. Typically, researchers only see a random platform ID and otherwise anonymized participant responses.

Despite anonymity on the research platform, modern data collection can inadvertently compromise participant privacy in subtle ways. Survey platforms often automatically gather metadata such as people's geographic location, which becomes a problem when combined with other survey responses. For instance, a participant from a small town who provides specific demographic or professional information may become identifiable (Morehouse et al., 2023; Wicherts et al., 2022). This risk increases when participants share sensitive information about personal experiences or behaviors.

To maintain participant privacy, researchers should take three steps. First, determine what participant information is truly necessary for the research objectives and ask only for that information. Second, create a plan for secure data storage. And third, establish clear protocols for data sharing that protect participant's identities. For example, many researchers will delete a participant's unique platform ID, geographic information, and any unusual demographic characteristics that might be used to identify who a participant is.

Other General Best Practices for Online Studies

While each section above addresses a specific ethical issue, the following sections outline five general practices that help researchers conduct more effective studies. Each section below is intended to help you translate ethical principles into effective research procedures.

Set Proper Expectations

When conducting an online study, researchers should provide sufficient information about the study before participants commit their time and effort.

Essential information to communicate upfront includes any special requirements or potentially concerning elements. Participants need to know if they will be asked to download software, participate in follow-up surveys, interact with other participants, make audio or video recordings, view sensitive content, or share personal information. While studies using deception may necessarily limit some disclosures, researchers should still communicate all requirements that might affect someone's willingness to engage with the project.

This transparency prevents participants from investing time in a study only to encounter unexpected requirements or other content they find objectionable later. Such situations not only waste participant time but can also compromise data quality.

As described in Chapter 14, it is a good idea to monitor the dropout rate when launching a study. Whenever the dropout rate exceeds 20 or 30 % it is an indication that something is surprising or upsetting participants.

Pilot Test Materials

Pilot testing study materials is an invaluable way to identify potential problems before they impact an entire study. Through pilot testing, researchers gain valuable insights into how participants interpret study materials and engage with experimental manipulations.

Beyond improving study materials, pilot testing provides practical benefits as well. A pilot study allows researchers to accurately gauge a study's duration, helping them set appropriate compensation rates. Technical issues such as broken links or malfunctioning survey elements can be identified and resolved. And, perhaps most valuably, pilot testing generates candid participant feedback about the research experience, enabling researchers to optimize their protocols before investing in full-scale data collection.

Always consider running the study with a handful of people before launching the full study.

Use Demographic Screening Appropriately

When studies require specific demographic criteria, researchers must carefully consider how to identify qualified participants without creating inefficient or frustrating experiences. While participant platforms often maintain demographic profiles that enable precise targeting, researchers frequently encounter situations that require additional screening or an expedited process. Some modern research platforms prohibit uncompensated screening questions at the start of a study while others allow it as long as participants are paid for their time.

Build Trust with Effective Communication

Clear communication between researchers and participants helps maintain a productive research environment. Beyond handling routine matters like technical issues or payment questions,

thoughtful communication helps create an atmosphere of mutual respect and professionalism in online research.

Researchers should maintain open channels of communication for several reasons. First, addressing technical problems or clarifying study requirements ensures smooth data collection. Second, providing clear explanations for rejections or bonus payments helps participants understand how their work is evaluated, as we saw in the opening story of this chapter. Third, sending timely reminders for longitudinal studies and follow-up tasks supports continued participation. Even simple acknowledgments or notes of thanks for people's contributions to a study can significantly enhance the research experience.

But communication between researchers and participants carries importance beyond its immediate purpose. In an environment where face-to-face contact is not possible, each communication represents an opportunity to build trust and demonstrate respect for participants. A brief message explaining a technical issue or acknowledging receipt of quality data can transform an impersonal transaction into a more meaningful one.

Issue Payments Promptly

After participants complete a project, researchers must decide who to approve and who to reject. During this time, participants wait. Establishing a systematic approach to reviewing submissions and processing payments can improve participant trust, the platform's reputation, and the health of the research ecosystem.

Most platforms require researchers to make payment decisions within two weeks of a participant's submission. However, waiting until the deadline creates uncertainty for participants, some of whom may be eagerly waiting for the payment. As a result, it is helpful for researchers to review submissions and process payments as quickly as possible after a study is completed.

Conducting Ethical Research: Writing an IRB Application

Module 15.2

Examine a sample IRB application and learn how to complete your own research proposal

In Module 15.1, we explored the ethical principles that guide research involving human participants. Now, we transition to the practical application of these principles, focusing on how to navigate the IRB application process.

Submitting an IRB proposal is a critical step for nearly all research conducted at academic institutions and other research organizations. This module aims to demystify the process. We will use an example research study—one investigating the relationship between eating, sleep, and depression to illustrate how to address the typical questions IRB boards ask. By walking through how one might answer these questions for a specific online study, you will gain a clearer understanding of how to prepare your own IRB applications. The specific answers to these questions can be used as a template for your own proposals.

The Purpose of an IRB Review

Before we examine specific IRB questions, recall that the purpose of an IRB is to protect the rights and welfare of human research participants. The IRB reviews your research plan to make sure it aligns with the ethical principles of Respect for Persons, Beneficence, and Justice, and complies with relevant regulations.

Common Components of an IRB Proposal: An Example Study on Eating, Sleep, and Depression

Let's imagine we are preparing an IRB application for a study that aims to explore associations between when people eat meals, when they sleep, and symptoms of depression. The study will be conducted through an anonymous online survey on the Connect platform and administered on Qualtrics. We will look at key sections of an IRB application and discuss how to form responses to different questions an IRB might ask.

Most IRB applications start by asking for general information about the project and the people conducting it. The application will often ask you to describe the study aims, background, and

significance. Following these questions the IRB is likely to ask you to describe the research design and methodology.

Research Design and Methodology

A typical IRB question will ask you to: "Describe your research design (e.g., anonymous online survey, experiment, qualitative interviews). What procedures will participants undergo step-by-step? What instruments, questionnaires, or tasks will be used?"

Your goal here is to give the IRB a clear picture of what a participant will do and experience from start to finish. Be precise and provide all necessary study materials as appendices.

Sample response

"This study will employ a cross-sectional research design using an anonymous online survey. The survey will be created and hosted on the Qualtrics platform.

Procedures: Potential participants will be recruited from the CloudResearch Connect online participant recruitment platform. They will first see an advertisement for a study titled 'Survey on Behavior.' This advertisement will specify the study's approximate duration (18 minutes) and compensation (\$2.50). Upon clicking to participate, people will be directed to the Qualtrics survey. The first page will be an informed consent document detailing the study's purpose, procedures, risks, benefits, confidentiality measures, voluntary nature of participation, right to withdraw, and researcher contact information (see Appendix A: Consent Form).

Participants will indicate their consent by clicking a button to proceed. They will then complete several validated questionnaires assessing their sleep quality, chronotype (morningness/eveningness preference), eating timing, attitudes toward eating, bingeing behaviors, and depressive symptoms. Finally, participants will answer a set of standard demographic questions. No deception is involved in this study. Upon completion, participants will be provided with a completion code to receive their compensation via the Connect platform.

Instruments/Measures: The survey will include the following validated measures: the PROMIS Sleep Disturbance Short Form (Yu et al., 2012), the Munich Chronotype Questionnaire (Roenneberg et al., 2003), the Morningness-Eveningness Questionnaire (5-item version; Adan & Almiral, 1991), the Food Timing Screener (Chakradeo et al., 2022), the Eating Beliefs

Questionnaire (EBQ-18; Burton et al., 2018), the Questionnaire on Eating and Weight Patterns-5 (QEWP-5; Yanovski et al., 2015), and the Center for Epidemiologic Studies Depression Scale-Revised (CES-D-R; Eaton et al., 2004). Demographic questions will include age, gender, ethnicity, and education level. Copies of all questionnaires are provided in Appendix C."

Subject Population

A typical IRB question will ask: "Who are your target participants? What are the inclusion and exclusion criteria? How many participants do you plan to recruit, and how did you determine this sample size? Will you be including any vulnerable populations?"

In answering these questions, be specific about who can and cannot participate and why. Justify your sample size and clearly state if vulnerable populations are involved (and if so, what extra protections are in place). Online participants are not considered a vulnerable population except in rare circumstances, and it is okay to state that explicitly in the application.

Sample response

"The target participants for this study are adults aged 18 years and older who are registered members of the Connect online research platform.

Inclusion Criteria: Participants must be 18 years of age or older and able to read and understand English, as the survey is administered in English.

Exclusion Criteria: Individuals under the age of 18 will be excluded. This study will not specifically recruit or include known vulnerable populations such as prisoners, pregnant women, or people with known cognitive impairments that would preclude informed consent. The online platform (Connect) requires users to be 18 or older.

Sample Size: We aim to recruit 1,250 participants. This sample size was primarily determined by looking at past studies that have aimed to study relationships between the variables of interest.

Recruitment Methods

A typical IRB question will ask: "How will potential participants be identified and recruited? What will recruitment materials (e.g., flyers, online advertisements, email scripts) say?"

Being detail-oriented here is key. The IRB wants to see exactly what potential participants will see. Many IRB members may not be familiar with the specific platform being used and so it is helpful to provide details to make this clear.

Sample response

"Participants will be recruited from Connect (connect.cloudresearch.com), an online platform where people voluntarily register to participate in research studies. Our study will be posted on the Connect participant dashboard, visible to registered users who meet basic eligibility (e.g., U.S. residents, 18+).

Recruitment Material: The study will be advertised on the Connect dashboard with the following information (it is helpful to include a screenshot of what this looks like, when possible):

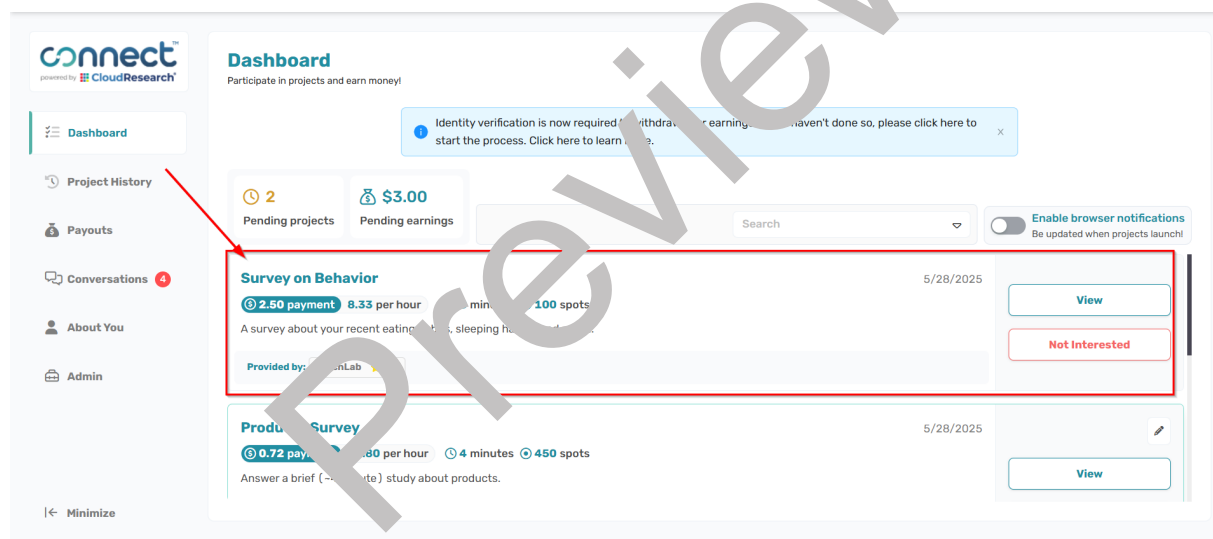


Figure 15.2. A sample study posting on the Connect platform.

People who are interested in the study can then click to view more details and, if they choose, proceed to the informed consent page and the survey itself. No direct emails or personalized solicitations will be used for initial recruitment beyond the platform posting.

Informed Consent Process

A typical IRB question will ask: "How will informed consent be obtained and documented? (Attach the consent form/script, see below). How will you ensure participants understand the information? How will you ensure participation is voluntary?"

This section is important as it directly relates to the principle of Respect for Persons. Your consent form is the primary document here. The IRB will want to see the entire form and often will even provide a template or specific sections of text they think should be included. (See the full consent form below). The IRB's primary concern with consent is to be certain that participants are making a truly informed decision about whether to participate.

For most online studies, you will need to request a waiver of documentation for informed consent. The reason for the waiver is simple: in online studies with platforms like Connect participants are remote and anonymous. Requiring people to sign a consent form would not only be impractical, but it would be the only thing that identifies who participated in the study. For these reasons, most IRB's grant a waiver of documentation for the consent process.

Sample response

"Informed consent will be obtained via an online information sheet presented as the first page of the Qualtrics survey. Participants must indicate their consent to participate before proceeding to the survey questions. This is a 'click-wrap' consent method commonly used and accepted for minimal risk online research where a signed document is not feasible and could compromise anonymity. We are requesting a waiver of signed consent for this study."

An example consent form is included below.

Risks and Protections

A typical IRB question will ask: "What are all potential risks to participants (even if minimal)? How will these risks be minimized? How will participant privacy and data confidentiality be maintained?"

IRBs need assurance that you have thought through any potential negative impacts, however small, and how you will protect participants. For online surveys, psychological risks (like

boredom or discomfort with questions) and privacy/confidentiality are typically minor but should still be explicitly stated on the IRB form.

Sample response

"Potential Risks: This study is considered to pose minimal risk to participants. Potential risks are no greater than those ordinarily encountered in daily life or during the completion of standard psychological questionnaires. These may include possible boredom or fatigue due to the length of the survey (approximately 18 minutes), minimal psychological discomfort when answering questions about mood, eating habits, or sleep patterns, as these can be personal topics.

Protections and Risk Mitigation:

Voluntary Participation & Withdrawal: The informed consent process will clearly state that participation is voluntary and participants can withdraw at any time without penalty by closing their browser window. They will also be informed they can skip any questions they prefer not to answer.

Anonymity/Confidentiality: No direct personally identifiable information (e.g., name, email address, specific date of birth) will be collected as part of the survey responses. Participants will be identified on the Connect platform by an anonymous alphanumeric string, which is used solely for awarding compensation and cannot be linked to their survey data by the researchers in any identifiable way in publicly shared datasets.

IP addresses collected by Qualtrics (if any, depending on settings) will be anonymized by the platform and from the dataset before any public sharing.

Data Storage: All electronic survey data will be stored on a password-protected computer accessible only to the Principal Investigator and the Faculty Advisor.

Data Retention and Sharing: The original de-identified dataset will be retained on the secure computer for a minimum of 3 years after study completion. After completion of the study and any publications, a modified, fully de-identified dataset, with any platform-specific alphanumeric identifiers and IP addresses removed, will be prepared and may be uploaded to the Open Science Framework (OSF), an online data repository, to promote scientific transparency and allow for

secondary analysis by other researchers, in line with open science practices. No information that could reasonably identify an individual will be included in any publicly shared dataset."

Minimal Discomfort: Questions are standard and widely used. If any question causes discomfort, participants are free to skip it. The study does not involve deception or exposure to distressing stimuli beyond what might be encountered when reflecting on common life experiences.

Contact Information: Participants will be provided with contact information for the researchers and the IRB should they have any concerns or experience distress."

Benefits

A Typical IRB Question will ask: "What are the potential direct benefits to participants? What are the potential benefits to society or scientific knowledge?"

Be honest and realistic. It is common for survey participants to receive no direct personal benefits beyond compensation. Focus on the scientific or societal value.

Sample response

"Direct Benefits to Participants: Participants are not expected to receive any direct benefits from participating in this research other than the monetary compensation provided for their time and effort. Some participants may find reflecting on their habits and mood to be interesting or informative, but this is not a guaranteed benefit.

Benefits to Society/Scientific Knowledge: The primary benefits of this study are scientific. The findings may contribute to a better understanding of the complex relationships between eating patterns, sleep quality, chronotype, and depressive symptoms. This knowledge could potentially inform the development of future public health recommendations or interventions aimed at improving mood and well-being by addressing modifiable lifestyle factors related to circadian rhythms."

Compensation/Payment

A typical IRB question may ask: "Will participants be compensated for their time and effort? If so, how much, in what form (e.g., money, gift card, course credit), and when will they receive it?"

Clearly state the compensation. Explain that for online studies compensation is typically delivered via the platform. Justify the amount as fair for the estimated time, referencing platform norms if possible. It is important to distinguish between platforms like Connect, MTurk, or Prolific where payment amounts are set by the researchers and market research panels where the researcher has no control over how much participants get paid. Below, we provide template answers for both types of data collection options.

Sample response for researcher-centric platforms:

"Yes, participants will be compensated for their participation. Upon completion of the 18-minute survey, participants will receive \$2.50. This payment will be processed through the Connect online research platform directly to their Connect account. This compensation rate is approximately \$8.33 per hour, which is consistent with current payment norms for similar online survey studies conducted on this platform. The payment is for completing the study; partial payment is not offered if the study is not completed, though participants can withdraw at any time."

Sample response for market research platforms:

"Yes, participants will be compensated for their participation. Upon completion of the 18-minute survey, participants will receive compensation in the amount that they have agreed to with the platform through which they enter the survey. While using this platform, we are not able to set the compensation rate. Most participants receive gift cards, rewards points, or cash equivalent to a few dollars or less. The payment will be commensurate with norms on the platform and for completing the entire study; partial payment is not offered if the study is not completed, though participants can withdraw at any time."

Example Consent Form

Below is a typical consent form that is included on the first page of the Qualtrics study.

This will appear at the start of the survey. Respondents will not be able to continue with the survey unless they click the arrow, indicating that they have read and understood the information.

Purpose

The purpose of this study is to explore the associations between eating, sleep, and mood.

Participation

It is estimated that 1,250 people will participate in this study. You must be 18 years or older to participate. If you decide to participate, you will be asked a number of questions about your recent eating habits, sleeping, and mood. Your participation is voluntary. You do not have to participate in this study. You have the right to withdraw from this study any time before you submit your responses by closing the survey window. There will be no penalty for exiting the survey, however you will only receive compensation upon completion of the survey. You may skip any questions that you do not wish to answer. The survey is expected to take approximately 15 minutes.

Risks

This project involves minimal risk and you should not experience any discomfort beyond what would occur in your everyday life in completing a standard medical or psychological evaluation. No information regarding your responses to the survey will be connected to you. To enhance your privacy and confidentiality, please complete this online survey on a personal computer or mobile device. If you use a work-issued computer or device, your confidentiality may not be protected.

Benefits

There is no direct benefit to you from participating however, it is hoped the knowledge gained will be of benefit to others in the future and may help inform evidence-based interventions for mood disorders.

Reimbursement for Participation

After the completion of the survey, you will receive \$2.50 through the Connect platform.

Confidentiality

All information collected in this survey will not identify you to the research team and cannot be linked to you in any way. If you would like documentation of your participation, please print out a copy of this consent page and the last page verifying the submission of your survey to the site. All data will be collected and stored for purposes of research only.

Contact Information

If you have any questions about the study or experience any study-related risk or discomfort, please contact _____. If you have any questions regarding your rights as a subject in this study, please contact the State University IRB Administration at _____ or via email at _____. By clicking to continue below, you acknowledge that you have read and understood the information above.

Summary

This chapter has explored the practical application of research ethics in online studies. While all human subjects research rests on the foundation of the three core principles established by the Belmont Report—respect for persons, beneficence, and justice—online studies face some special challenges in ensuring these practices are met. While most online studies present minimal risk to participants, researchers must carefully consider how these principles translate into concrete practices.

When participants report distress in online studies, it typically stems more from aspects of finding studies and interacting with researchers and less from the content of studies themselves. To improve participants' experience and uphold the practices of ethical research, you can: design user friendly studies, communicate clearly with participants, ensure data privacy and security, pay fairly, and offer access to support resources, when necessary.

Before conducting an online study, the IRB application process serves as a crucial checkpoint for ensuring these ethical principles are embedded in research. By systematically addressing questions about methodology, participant protection, risks and benefits, and data management,

researchers demonstrate their commitment to ethical conduct while creating a roadmap for responsible research.

In the next chapter, we will examine how to write about online research. In particular, we will describe how to write a Method section that adequately describes your source of participants and sampling strategy. Then, we will provide some general advice about writing scientific papers.

Additional Readings

- Litman, L., & Robinson, J. (2020). Conducting ethical online research: A data-driven approach. In L. Litman and J. Robinson (Eds.) *Conducting Online Research on Amazon Mechanical Turk and Beyond* (234-263). Sage Academic Publishing. Thousand Oaks: CA
- Moss, A. J., Rosenzweig, C., Robinson, J., Jaffe, S. N., & Litman, L. (2023). Is it ethical to use Mechanical Turk for behavioral research? Relevant data from a representative survey of MTurk participants and wages. *Behavior Research Methods*, 55(8), 4048-4067.
- Fowler, C., Jiao, J., & Pitts, M. (2023). Frustration and ennui among Amazon MTurk workers. *Behavior Research Methods*, 55(6), 3009-3025.

Chapter 16: Writing

Science is built on the idea that people will share their work. Sometimes that happens in a conference talk or an invited lecture. In some cases, scientists share their work in a fancy TEDTalk, a book, or an address to policy makers. But the most common way to share scientific research is in a humble journal article. Why do scientists write these?

In the book *Write it Up*, psychologist Paul Silvia offers several reasons for publishing academic research. At the top of his list are noble reasons: to share knowledge, advance scientific understanding, and have a positive impact on the world (Silvia, 2014). Not all writing meets those lofty objectives, though. People write for practical reasons like to get a job, complete a dissertation, or fulfill the requirements of an assignment. And because nature being what it is, people sometimes write for unsavory reasons: vanity, spite, to maintain a reputation.

As a student, your reasons for writing will often be practical: you need to complete an assignment, write a thesis, or submit a manuscript. However, to these practical reasons you can add the goal of learning; each writing assignment offers the chance to develop your knowledge and skill as a writer.

Even though the principles of good writing stretch across genres, scientific writing often follows certain conventions. Within the behavioral sciences, these conventions come from the American Psychological Association's (APA) publication manual. APA style papers often have an Introduction, Method, Results, and Discussion sections (Figure 16.1). Within each section, readers expect certain pieces of information.

In a Method section, for instance, readers expect to find information about who served as participants, how they were recruited, what they did, how the researchers chose the measures participants completed, and anything else relevant to assessing the appropriateness of the sample and measures for addressing the research question.

The Title of the Paper is Repeated at the Top of the First Page: Centered and Bolded

The introduction begins right away, without a heading of its own. This is where you give background and lead up to your hypotheses. The introduction usually has subsections.

The First Section Heading is Centered

Level 2 Headings are Left Aligned

Level 3 Headings are Italicized

Method

The method section uses a Level 1 header, and typically has several subsections, the most common are listed below.

Participants

Design

Material

Scale Name

Procedure

Results

Discussion

References

Figure 16.1. Papers written in APA style follow a formula. From the large section headings to the information communicated within each section all the way down to how the headings align throughout the paper. Following this format facilitates clear communication among scientists.

The goal of this chapter is to show you how to write an effective Method section. Because online research presents challenges that offline research does not, we highlight some unique details that should be a part of your sample description when your data are gathered online. We also explain features that should be a part of all Method sections and how to ensure your Method is complete. At the end of the chapter, we provide a sample from a published paper and highlight how it

meets the standards of article reporting detailed by the APA (American Psychological Association, 2020). Finally, we provide some general tips about writing at the end of the chapter.

Preview

Chapter 16 Outline: Writing

Module 16.1: Writing a Method Section



Reading

Learn about the pieces of a method section and how to describe online studies

- Understand the purpose and structure of Method sections including essential subsections like Participants, Procedure, and Measures
- Learn to write effective descriptions of participants, procedures, and measures with appropriate detail for online studies
- Apply open science practices including data sharing statements, preregistration disclosure, and transparency in reporting

Module 16.2: Writing Advice



Reading

Explore what makes writing good and how to improve your own writing

- Learn principles of effective scientific communication including writing with clarity, simplicity, and engaging openings that grab reader attention
- Practice techniques for clear scientific writing such as avoiding jargon, limiting nominalization, and using active voice with subject-verb-object structure
- Develop skills for structuring scientific arguments by focusing on ideas rather than studies and creating strong conclusions that reinforce key findings

Writing a Method Section

Module 16.1

Learn about the pieces of a method section and how to describe online studies

The Purpose of the ‘Method’ Section

Each section of a scientific paper has a purpose. The Introduction frames your research question and why it’s important; the Method tells readers what you studied, how you studied it, and who participated in your project; the Results describe what you found; and the Discussion summarizes what it all means. Even though the Method is sandwiched in the middle, many people write it first. That is because the Method is often straightforward: you tell people what you did and how you did it.

While description is the practical goal of a Method, there is also a rhetorical purpose. You want to convince readers that your methods were effective and fit for addressing your research question. To do that, you need to provide them with information to evaluate both the reliability and validity of your results. There is a catch, however: you have limited space.

Most articles have word limits. And, even if they don’t, readers have attention limits. Writing an effective Method requires discernment. You must decide what information goes in, what gets omitted, and what belongs in an Appendix or online Supplement. You must also write this section well enough to engage your readers in the minutia of your methods.

Parts of a Method Section

In the book *Scientific Writing for Psychology*, Robert Kail (2015) says you should “use subheadings freely” in the Method section.

Some subheadings are necessary. Everyone, for instance, expects a section describing “Participants and Design” or “Measures and Outcomes.” But beyond that your paper may contain sections describing the “Procedure,” “Apparatus,” or “Open Science, Data Sharing, and Transparency” practices.

The specific subheadings you choose should reflect the unique elements of your research. For experimental studies, consider separate subsections for Design, Procedures, and Materials. For

online studies, an additional subsection addressing Data Quality Measures may be warranted. When determining which subheadings to include, consider what information readers need to replicate your methodology, evaluate the appropriateness of your methods, and understand any methodological limitations.

Let's look at some subsections in detail.

Participants and Design

The participants section describes who participated in your project and how they were sampled.

Where did you recruit participants? How many people took part? How many dropped out early? What were the demographic characteristics of the people who completed the project? When was the data gathered? How much were participants compensated? How long did the study take participants to complete? These are some of the questions a “Participants” section should answer.

For online studies specifically, you should also address:

- The platform used for recruitment (e.g., Connect Mechanical Turk, a university participant pool)
- Any screening criteria or attention checks used
- Geographic restrictions on participation
- Browser or device requirements
- Completion rates and patterns of attrition
- Steps taken to ensure data quality

The norms for describing sample demographics vary by discipline. The APA’s journal article reporting standards say to “Report major demographic characteristics...and important topic-specific characteristics (e.g., achievement level in studies on educational interventions)” (American Psychological Association, 2020). For most studies conducted online, researchers report the age, race, ethnicity, gender, and topic-specific characteristics of the sample. If you have a lot of demographic information, a table is a great way to present it. A table allows you to present a lot of information in a small space, as shown in Table 16.1.

Category	Sample		
	Standard	Open	Inexperienced
Annual Household Income			
< 20k	15	20.5	14.2
20-39k	28.5	27.1	20
40-59k	23	19.8	25.5
Marital Status			
Married	35	33	42
Divorced	10	9	9.5
Never married	55	58	48
Race			
White	70	77	82
Black	12	10	8
Asian	10	7	5
Highest Degree			
No college degree	42	35	38
College degree	50	55	52
Political Party			
Republican	27	20	21
Democrat	45	48	32
Independent	28	26	27
Religion			
Christian	40	42	50
Atheist	25	22	18
Other	10	8	10

Table 16.1. Demographic information from Robinson et al. (2019). Standard, Open, and Inexperienced refer to three different samples of participants from one online source.

Beyond describing the sample, a “Participants” subsection often includes information about the study design and how sample size was determined. Where appropriate, researchers report the statistical power they had to detect the effect of interest. If your project contains an experiment, near the end of this section is a good place to describe your design and how participants were assigned to conditions.

An example “Participants” subsection appears below. It comes from the same paper referenced above about online research participants (Robinson et al., 2019). As you will see, it covers many of the points outlined above. Other points, like the sample source and why it was appropriate, were explained before the Method section. The topic of the paper concerned how a participant’s level of experience on sites like Mechanical Turk may present a problem for some types of research.

Participants section from Robinson et al., (2019). “We aimed to collect data from 750 people—250 in each of our three samples (standard, open, inexperienced). We expected the study to take about 15 minutes and paid each participant \$1.00. Although we did not conduct a formal power analysis, we aimed to recruit large samples in line with past work examining data quality on MTurk.

The final dataset included 768 responses. There were more responses than participants we aimed to collect data from because two participants entered the study more than once and 34 people dropped out of the study early. We retained data from all participants who completed all our measures of data quality ($n = 750$). This cutoff resulted in removing incomplete responses from the two people who entered the study more than once and eight people who completed less than 30% of the study. After exclusions, the sample included nearly equal numbers of men ($n = 375$) and women ($n = 350$), and the average age was 33.9 years ($SD = 10.7$) (see Table 1 for detailed demographic information).”

Procedure

The procedure describes what happened in the study and, when done well, why things occurred as they did. Some procedure sections are long. Most are short.

Regardless of length, many researchers get the “Procedure” section wrong because they don’t include enough information. The Procedure should contain enough detail for readers to evaluate your methods and potentially replicate your study.

Topics traditionally covered in the Procedure include: the source of participants, approach to sampling, inclusion and exclusion criteria, and other details needed to evaluate the study. Within online studies, researchers should also report the compensation participants received, how long the study took, anything special about the sampling process, and any ethical considerations.

As an example of how omitting details impedes a reader's ability to evaluate the research, let's return to Mechanical Turk. Shortly after becoming a popular source of data (Buhrmester et al., 2011), most researchers started using MTurk's reputation metrics to restrict who could participate in their studies (e.g., Peer et al., 2014). However, reporting the use of these qualifications never caught on.

Failing to report information like the use of reputation metrics matters because it affects who can participate in the study. When researchers use MTurk's reputation qualifications, the newest and most inexperienced participants—about 35% of the population overall—are blocked from participating (because they do not have a reputation yet; see Robinson, et al., 2019). Thus, even a small decision focused on improving data quality can have a big influence on sampling, and these are exactly the kinds of decisions that should be disclosed in the Procedure section.

As another example of underreporting, consider “nationally representative” data. Across many behavioral science disciplines, nationally representative data are as we talked about in Chapter 9. Yet, over the years, several polling organizations and some participant recruitment companies have taken to calling data that were either weighted, gathered using non-probability quotas, or both, “nationally representative.” When behavioral scientists repeat this description in papers without detailing their sampling approach, readers have little chance of adequately understanding the research or its results.

Beyond sampling details, the procedure should inform readers about what participants did in the study. For short surveys or questionnaires, you may write something like, “Participants completed the following scales in a random order, with item order within each instrument also randomized. A full list of items and other related material can be found in the Supplementary Materials section” (Rivera et al., 2022). Following this sentence, another subsection may describe each scale in the study.

For longer or more complicated projects, the procedure section may be detailed. A daily diary study, longitudinal investigation, video interview, dyadic study in which participants interact with one another, or other similarly complex study would require more description than a short survey. In a longitudinal investigation, for instance, you should detail how you recontacted

participants, what attrition occurred between each wave, and what incentives were offered to increase retention.

Finally, if necessary, the Procedure is a good place to describe any agreements with your institutional review board, any unusual ethical standards the research adhered to, or any safety monitoring the study required.

As an example of what the Procedure section can look like, consider the paragraph below. It comes from the same paper as the Participant description section above (Robinson et al., 2019).

“To recruit participants, we created three separate studies on MTurk and varied the worker qualifications for each. All three studies were set up and managed using the TurkPrime platform [23]. In the first study (Standard), we used standard worker qualifications of at least a 95% approval rating and more than 100 HITs completed. In the second study (Open), we used no qualifications, meaning the study was open to all workers on MTurk. Finally, in the third study (Inexperienced), we required workers to be inexperienced by setting the qualification requirement to less than 50 HITs completed. Data collection for all three studies started at the same time and ended after approximately one day (standard = 20 hours, open = 22 hours, inexperienced = 25 hours). After all three studies ended, we used the TurkPrime database to query workers’ approval rating and number of HITs completed in the open sample.

Participants completed the Asian Disease experiment, Mt. Everest experiment, Trolley Dilemma experiment, Big Five Personality Inventory (BFI), Cognitive Reflection Test (CRT), and demographic questions. Each experimental manipulation—Asian Disease, Mt. Everest, Trolley Dilemma—had two conditions and participants were randomly assigned to conditions. The order of the experimental manipulations, the BFI, and the CRT was randomized across participants. After participants completed all tasks, they answered demographic questions. We included four attention check questions at various points in the survey—two in the BFI and two in the demographics section.”

Apparatus

The Apparatus section describes any novel methods, technologies, or tools used to collect data. While not all studies require this section, it becomes essential when implementing innovative approaches that readers may be unfamiliar with.

In online research, a new “Apparatus” is often some software, hardware (i.e., a device), website, or technique for collecting data. Whenever you use a new method or technique, you should give readers a detailed description and information about where they can learn more. Examples of new methods are using voice recordings to learn about participants’ reasoning processes (Ristow & Hernandez, 2023), using mobile devices to capture data about behavior (Harari et al., 2021), or even using passive recording technology to assess the quality of participants’ responses (Permut et al., 2019). In each of these papers, the researchers described the new method, explained how it works, and made a case for its validity.

Measures and Outcomes

This subsection is easy to get right. All you do is describe what you measured what you measured and make a case that your measures were appropriate (i.e., reliable and valid).

Just as with the Apparatus section, measures and outcomes (or sometimes just “Materials”) that are commonly used within your field require less explanation than things that are new. For well-established measures, all you need to do is describe the measure and provide a reference to it. You should tell readers how many items the measure has, what kind of response scale participants were given (e.g., 1 to 7, -3 to 3), what the scale labels or anchors were, and what evidence there is for the measure’s reliability and validity.

For new measures, you must spend more time describing why the measure was appropriate for the situation and outlining evidence of its psychometric properties. Even if the evidence is preliminary and even if it’s based on your own data (as is likely to be the case), readers will want to know why your measure is suited to the situation. It’s your job to tell them.

We present the Measures subsection from a journal article that used online participants to investigate what motivates people to stay single as they move throughout life.

From Park et al., 2023:

Measures: Profile Indicators

Fundamental Social Motives. All participants completed eight subscales (each consisting of six items) from the Fundamental Social Motives Inventory (FMI; Neel et al., 2016). Note that among the 11 subscales of the FMI, three related to mate retention or parenting were not included in the survey as they were not relevant to (all) singles. The internal consistency was high for all subscales as follows: self-protection (e.g., “I think a lot about how to stay safe from dangerous people”; $\alpha = .88$), disease avoidance (e.g., “I avoid places and people that might carry diseases”; $\alpha = .88$), affiliation—group (e.g., “I enjoy working with a group to accomplish a goal”; $\alpha = .86$), affiliation—exclusion concern (e.g., “I would be extremely hurt if a friend excluded me”; $\alpha = .89$), affiliation— independence (e.g., “Having time alone is extremely important to me”; $\alpha = .84$), status (e.g., “It’s important to me that other people look up to me”; $\alpha = .81$), mate seeking (e.g., “I spend a lot of time thinking about ways to meet possible dating partners”; $\alpha = .93$), and kin care—family (e.g., “It is extremely important to me to have good relationships with my family members”; $\alpha = .92$). All items were assessed on a 7-point scale, ranging from 1 (strongly disagree) to 7 (strongly agree).

Measures: Predictors

Background Variables. Four variables assessed at background were examined as predictors of profile membership. These include gender (men vs. women), age, dating history (have vs. have not been in a relationship before), and marital history (ever vs. never been married). Note that given the limited number of individuals belonging to the “other” category for gender ($n = 6$), we dropped them from the analysis including gender. The number of divorced and widowed individuals was also small ($n = 30$), thus we collapsed ever-married individuals into one category. We kept these individuals in the model given previous work suggesting potential differences in never-married versus ever-married individuals’ social networks (Pinquart, 2003).

Attachment Insecurity. The Experiences in Close Relationships—Relationship Structures questionnaire (Fraley et al., 2011) was used to assess global (i.e., relationship-general) attachment insecurity. Participants responded to six items assessing attachment

avoidance (e.g., “I don’t feel comfortable opening up to others”; $\alpha = .86$) and three items assessing attachment anxiety (e.g., “I often worry that other people do not really care for me”; $\alpha = .88$) on a 7-point scale (1 = strongly disagree; 7 = strongly agree).

Fear of Being Single. Participants responded to the Fear of Being Single scale (Spielmann et al., 2013) which includes six items such as “I feel anxious when I think about being single forever” ($\alpha = .85$). The items were rated using a 5-point scale (1 = not at all true; 5 = very true).

Measures: Outcomes

Satisfaction With Being Single. The Satisfaction With Relationship Status Scale (Lehmann et al., 2015) was used to measure satisfaction with being single. Participants were asked to think about their current relationship status (which, for all the participants, would be being single) and respond to questions such as “How happy are you with your current status?” ($\alpha = .92$) using a 4-point scale (1 = not at all; 4 = to a great extent).

Life Satisfaction. Participants responded to the Satisfaction With Life Scale (Diener et al., 1985) using a 7-point scale (1 = strongly disagree; 7 = strongly agree). Items include five statements such as “In most ways my life is close to my ideal” ($\alpha = .89$). Correlations among all study variables can be found in the Supplemental Material.

Reporting Data Cleaning

An important part of the Measures and Outcomes subsection is explaining how you cleaned and screened the data. These measures should ideally be pre-registered (see Box 11.1 for more about pre-registration). Within the paper, however, you need to explain how you picked these measures, what evidence there is for their effectiveness, and how people performed on them.

If you followed the advice we gave for using attention check questions in Chapter 7, your Measures section should include a heading that says “Data Screening.” Within this section, you should describe the instructed response, nonsense, or nearly non-existent event items you used to screen for data quality. You should also mention any open-ended items you used to measure quality and what criteria constitute a passing or failing answer. Finally, you should report how many participants were excluded based on screening.

Open Science, Data Sharing, and Transparency

Science has changed a lot in the era of online research. Some changes are thanks to technology (e.g., Anderson et al., 2019; Buhrmester et al., 2018) and some are the result of shifts within the scientific community. One large shift is in the planning, reporting, and archiving practices of behavioral scientists, a change broadly known as “open science.”

Many of the methodological changes that fall under the umbrella of open science are intended to make research more transparent, more cumulative, and more collaborative. These changes have been spurred by fraud, shoddy research practices, and an inability to replicate the published findings from various disciplines (Bhattacharjee, 2013; Klein et al., 2018; Simmons et al., 2011). In a sign of progress, many researchers have changed the way they conduct and report their studies (Nosek et al., 2022).

What this means for the Method section is that many readers will expect to see a description of open science practices. These statements typically include information about how the sample size was determined, which cases of data were removed from analyses, and whether all manipulations and measures are being reported in the study (and if not, why not; Simmons et al., 2012). Open practice statements also typically include information about whether the study was pre-registered (along with a link to the pre-registration), how interested parties can obtain the data files, study materials, and analysis code (usually with a link to the repository), and whether the authors have any conflicts of interest. Sometimes, researchers add other disclosures such as their position toward the data or research question (e.g., Ledgerwood et al., 2023), but overall, the aim of these statements is to make the reporting of research findings more clear and more open.

Statements about open science are often short. Because manuscripts typically have word limits, the disclosures researchers provide are sometimes scattered throughout the Method section.

Whenever necessary, they can also be placed under their own heading labeled something like “Open Practices Statement” or “Data Sharing.” Two examples of these kinds of statements are below, reflecting the variability these statements can entail:

Data Sharing Statement - From Clifton & Kerry, 2023

“Data, study materials (including all measures administered but not relevant to this study), and code are publicly available:

(https://osf.io/r3ksa/?view_only=f240306aede3473a8551729a6fb9bf34).”

Ethics and Open Practices Statement - From Sun et al., 2022

“We used data from three of our existing datasets. Data collection and coding procedures for Sample 1 were approved by Institutional Review Boards (IRBs) at Washington University in St. Louis (IRB ID: 201206090; Study Title: Personality and Intimate Relationships Study) and the University of California, Davis (IRB ID: 669518–15; Study Title: Personality and Interpersonal Roles Study). Data collection procedures for Samples 2 and 3 were approved by the IRBs at the University of Pennsylvania (Sample 2; IRB ID: 831767; Study Title: Moral Change Goals) and the University of California, Davis (Sample 3; IRB ID: 1328211-2). Data collection procedures for trait ratings (which we use for supplemental analyses; see Supplemental Materials, Sections 6–7) were approved by the IRB at University of Pennsylvania (IRB ID: 844199; Study Title: Best and Worst Trait Ratings).

For Sample 1, we used data from the first wave of the longitudinal Personality and Interpersonal Roles Study (PAIRS). Other published articles have used the PAIRS dataset (for a full list of citations, see <https://osf.io/3uag4/wiki/home/>). A few articles used the self- and informant-reports of personality traits that we use in supplemental analyses, but none have used the best and worst trait measures included in this study. For Samples 2 and 3, we used data from a study on personality change goals. The previously published article using these samples (Sun & Goodwin, 2020) used the self- and informant-reports of personality traits that we use in supplemental analyses, but did not use the best and worst trait measures. Codebooks for all measures in these datasets are available at <https://osf.io/jce7k/>. Below, we describe the measures and procedures relevant to the current article.

The codebook, data (posted in a way that prevents targets from finding out what their friends said about them), and R scripts required to reproduce the analyses reported in this paper are available at <https://osf.io/jce7k/>. We did not preregister these analyses as we

were already familiar with the datasets when we conceptualized this project. Instead, to limit the risk of overinterpreting potentially spurious effects, we highlight the findings that replicate across at least two samples (at a conventional $p < .05$ threshold) and are therefore more likely to be robust.

The effects reported in the results section that met this replication threshold also met an alternative standard of evidence for claims of new discoveries—whether the effects are significant at a $p < .005$ threshold (v et al., 2018)—in at least one sample. Note that we coded and analyzed a few additional variables in the Sample 1 data for an undergraduate research project (see Supplemental Material, Section 1). We later refined the scope of the current paper to the variables that are presented in this paper and coded only these variables in the Sample 2 and 3 data. Apart from the additional variables coded in Sample 1, we report all coded variables.”

Once your Method section is complete, it’s time to turn to the Introduction, Results, and Discussion sections. There are many great resources for learning to write these sections effectively. Below, we present a complete Method section from a recent paper with the hope that you see how it meets the criteria presented throughout this chapter. After presenting this method, we share some general advice for writing that you can use throughout your manuscript.

Example Method from Litman et al., 2023:

Method

Participants and design

Study 1a Three hundred and two adults from MTurk participated in Study 1a. We used CloudResearch’s MTurk Toolkit (Litman et al., 2017) to target participants within the United States and to recruit participants in different age groups. Specifically, we recruited 50 participants in six separate groups, with each group corresponding to a different decade of age (20s through 70s). Participants were paid \$0.50 to complete the study which we estimated would take 3 minutes. All data were collected in April 2019, and data collection ended after 3 days.

Study 1b We recruited 350 adults from Prime Panels. As with Study 1a, we split the sample into six groups of approximately 50 participants each, with each group corresponding to a different

decade of age. Because Prime Panels aggregates several panels to collect large samples, participants were compensated based on the platform they were recruited through. Some participants may have completed the study in exchange for flight miles, points, money, or other rewards. All data were gathered in April 2019; data collection closed after 3 hours.

Procedure

We presented participants with the AVI. Some of the questions had four response options and some had five. We instructed participants to answer to the best of their ability without using outside sources and to select “I don’t know” when applicable. We also stressed that we would not penalize participants if they did not know the answers. As in the instrument development study, we asked participants to provide information about their age to verify that the database information was accurate. These included open-ended questions about participants’ current age, the year they graduated from high school, and how old they were during Watergate (participants under 50 typically wrote “not born yet”). For exploratory purposes we also asked participants to select the decade of their life in which they were the happiest and to elaborate on what was positive about that time. They then selected the decade of their life that was most difficult and described what made it so. The results from these questions are not reported here.

Analytic approach

We used the difference method to assess each person’s relative knowledge of historical (questions about pop culture prior to the year 2000) and contemporary (questions about pop culture after the year 2000) culture. For each participant, we separately summed the number of correct responses on all items measuring historical and contemporary knowledge, then converted the sums to percentages. Finally, we calculated a difference score by subtracting the percentage of correct responses to contemporary questions from the percentage of correct responses to the historical questions. This yielded a difference score variable with a range of -100% (correctly answered all contemporary questions and no historical questions) to $+100\%$ (correctly answered all historical questions and no contemporary questions). For Study 1a, in addition to using the CloudResearch database to target participants whom we expected to fall into six age groups, we asked participants to self-report their age. We opted to rely on self-reported age in our analyses because that is the data most researchers would have access to. There was a strong

correspondence between self-reported age and database age ($r = .965, p < .001$). In both studies we used linear regression, predicting the continuous self-reported age variable using performance on the AVI items as the predictor. We also assessed the utility of the instrument for distinguishing between decades of age. To do so, we split participants' self-reported age into six groups, with each group corresponding to a different decade, and tested differences between the groups using a one-way ANOVA. To assess the value of using the difference score, as opposed to just using the scores on the contemporary or historical questions, we also examined the correlations between self-reported age and each of these three measures. Here and in later studies we tested both the full 19-item instrument (see Table S1) as well as a shorter six-item subset (AVI-S). In addition to being easier to implement, research indicates that much of the age-related differences in people's knowledge can often be captured by a few items rather than by multiple items (e.g., Schroeders et al., 2021). The six-item subset comprised the three historical items that older adults most often answered correctly and the three contemporary items that younger adults most often answered correctly (see Table 1). Historical items include Bonanza (1959–1973), The Way We Were (1974), and The First Time I Ever Saw Your Face (1969), while contemporary items include Somebody That I Used to Know (2001), How You Remind Me (2001), and Boom Boom Pow (2009). Most analyses showed the difference between the full 19-item scale and the shorter six-item version was insubstantial (e.g., the full scale predicted 68.2% of the variance in age, while the shorter version predicted 67.2%). Therefore, we report the results of the shorter scale. Analyses using the full version of the AVI are available in the supplemental materials.

Writing Advice

Module 16.2

Explore what makes writing good and how to improve your own writing

Beyond accurately describing your methods and reporting your results, there are some lessons you can learn to ensure people will want to read what you write. We have gathered some of those lessons here.

The Transaction

When setting out to write a scientific paper, it's important to remember that scientific writing, like all writing, is an act of communication. The writer attempts to put their thoughts and some part of themselves into words. This transaction works best when people don't like, well, people.

Too often, people feel pressured to adopt a formal or scholarly tone when they write about scientific topics. This is a mistake. The goal of writing is not to impress readers with jargon or convoluted sentences, but to convey your research clearly and effectively. If you're passionate about the topic and you write well, readers will be carried along in your wake.

Effective scientific writing bridges the gap between your study's findings and your audience's understanding of the issue in question. For success, your aim should be to express your thoughts and discoveries in a manner that is accessible, precise, and engaging. The best way to do this is to embrace simplicity, keep the reader's perspective in mind, and know that a little warmth and a little humanity go a long way in scientific writing.

The Opening

What is the purpose of a paper's first sentence? Answer: to entice the reader into the second sentence!

Nearly every book about writing contains the admonition to "Start strong." This advice is often repeated for a reason: readers who are unenthused by the first sentence seldom make it to the second. That means your opening, or lede, needs to grab the reader and convince them that your article is worth reading.

What are ways to grab attention? A surprisingly simple method is to ask a question. For example, here's how two researchers began a paper on a cognitive bias known as the anchoring and adjustment heuristic:

“In what year was George Washington elected president? What is the freezing point of vodka? Few people know the answers to these questions, but most can arrive at a reasonable estimate by tinkering with a value they know is wrong” (Epley & Gilovich, 2001).”

With just two questions and one sentence, the reader is drawn into the paper.

Another effective strategy is to begin with an interesting statistic, fact, or observation. That is how a 2019 paper by Schroeder et al. investigating the role of handshaking in promoting deal-making began:

“After years of negotiations between Prime Minister Shinzo Abe of Japan and President Xi Jinping of China, diplomats from both countries managed for the two leaders of Asia's biggest economies to meet at a 2014 economic summit for a single purpose: to shake hands. The handshake took months of scheduling to arrange, with the news media noting that the “small gesture holds great importance” for future negotiations and would be “parsed for deeper meaning.” (Schroeder et al., 2019).”

Finally, a third strategy is to directly comment on behavior:

“When children draw on walls, reject daily baths, or leave the house wearing no pants and a tutu, caretakers may reasonably doubt their capacity for rational decision-making. However, recent evidence suggests that even very young children possess sophisticated decision-making capabilities for reasoning about physical causality (e.g., Gopnik et al., 2004, Gweon and Schulz, 2011), social behavior (e.g., Gergely, Bekkering, & Király, 2002), future events (e.g., Denison and Xu, 2010, Kidd et al., 2012, Téglás et al., 2011), concepts and categories (e.g., Piantadosi et al., 2012, Xu et al., 2009), and word meanings (e.g., Xu & Tenenbaum, 2007).” (Kidd et al., 2013).

Each of these methods is effective because they involve people thinking, feeling, and behaving in ways that behavioral scientists care about. There are, of course, other ways to begin a paper such

as telling a story (Neel & Lassetter, 2019), using a quote (Gray et al., 2014), or pointing out a problem or contradiction (Kerry et al., 2023). Regardless of how you open your paper, do it in a way that grabs the reader's interest and sets the stage for your research question.

Imitation

No one is born knowing how to write. Learning takes practice, practice, and more practice. Yet one way to accelerate this process is to imitate.

Imitating good writing doesn't mean you should try to sound like someone else; you have your own voice that only you can bring to the page. But imitation does mean you should try to examine what good writers do and do some of it in your writing.

The ways to open a paper above are examples of something that can be imitated. Everyone can try to pose an interesting rhetorical question about the topic of their research. Similarly, if you read good writing, you will notice other things that can be imitated such as simple word choice, how to frame an introduction or report results, and how to end a paper on a strong note. Many good writers learned what they know by imitating others. You should do the same.

Actions and Ideas

At the heart of good writing is a subject → verb → object structure. This structure works because it helps readers visualize who is doing what, and active verbs drive the writing forward.

Unfortunately, the subject → verb → object structure gets easily buried under ideas, especially in scientific writing. For example, consider this sentence: "Does a nuanced understanding of the normative interplay between architectural design elements and cultural connotations among seasoned professionals predict the subjective aesthetic evaluations of architectural compositions by novice observers?" Who is doing what in this sentence? What is the study about? It's hard to tell because the actions the researcher is interested in have been buried beneath ideas. A more direct way to say the same thing is: "Do people who know a lot about how buildings are designed have similar opinions about architecture as those who are new to it?"

What makes a lot of scientific writing hard to read is something the writer Helen Sword calls [nominalizations or zombie nouns](#) (Sword, 2012). Zombie nouns are words that used to be a verb or an adverb but someone, somewhere decided would work well as a noun. For example, the

word *expect* is a verb that often appears in print as “expectation.” The word *crony* is a noun that appears as *cronyism*. It isn’t necessarily a problem when one or two of these nominalizations appear in a sentence—they can help express complex ideas. The problem, however, is when too many nominalizations appear in the same sentence—then, they kill your writing. Sentences that are loaded with nominalizations bury your concepts and ideas in abstractions, preventing readers from following who is doing what.

Scientific writing leans toward nominalizations because writers need to express abstract ideas. If you remember Chapter 1, the variables that behavioral scientists study—life satisfaction, self-efficacy, happiness, regret—are not physical things but abstractions. As a result, it’s easy to get lost talking about the concepts within a study and forget that there are supposed to be people doing the thinking, feeling, and behaving behind the data points. Thus, limiting nominalizations is one key to clear writing.

Jargon

Science is full of jargon, and not all of it is bad. Jargon helps experts communicate complex ideas with each other (jargon is often a special case of the nominalizations we talked about above). But not all jargon is good. In fact, most of it is bad because it impedes clear communication.

The most common type of academic jargon is the nominalizations we described above, but a close second is acronyms. Some acronyms are useful. But when acronyms are unnecessary or overused, they can make writing hard to read.

For instance, if you are a social psychologist, you likely know that if someone starts talking to you about how WEIRD their sample of participants is, WEIRD has a specific meaning other than unusual or strange (it stands for Western, Educated, Industrialized, Rich, and Democratic). WEIRD is a useful acronym because it helps researchers remember that explanations for human behavior often need to apply beyond the small number of Western developed nations where most behavioral research has been conducted in the past (Henrich et al., 2010).

In many other instances, acronyms are not helpful. Many researchers, for instance, have probably found themselves reading a paper in which the authors created an acronym to refer to the

measure or scale used in the study, but the acronym is not intuitive or easy to remember. It may simplify the writer's task, but it makes more work for the reader. This is a bad acronym.

In many cases, the best policy will be to eliminate as much jargon as possible. This will also make your work accessible to people beyond your field or area of expertise.

Ideas First, (Scientists Second)

Scientific writing is part of an ongoing conversation. Your work should contribute to what has come before. Nowhere is this connection clearer than in the introduction.

In the Introduction to a scientific paper, your task is to describe the relevant research that motivated your study and set the stage for how your research will address the question you were interested in. In describing previous research, your writing will be better if you focus on what is and isn't known about your topic rather than on what past studies or researchers have shown. To see the difference, consider the examples below.

An Example with Scientists and Studies First. Kahneman (1985) conducted foundational work on counterfactual thinking. By asking participants to think about actions they have taken and actions they could have taken but didn't (inaction), he found that regrets of action tend to be stronger and more common than those of inaction because people often find it easier to imagine undoing an action they took (and mentally returning to the status quo) than to imagine what would have resulted from something they never did in the first place. However, research by Gilovich and Medvec (1994, 1995) shows a more complicated picture. Using surveys, interviews, and two experiments, they had participants think about both the short and long-term effects of both actions they took and regretted and things they never did but regretted not acting upon later. Their results indicated a temporal shift—namely, that people's regrets of action are more intense in the short term but regrets of inaction gain prominence in the long run.

An Example with Ideas and Findings First. Foundational work on counterfactual thinking indicated that regrets of action tend to be stronger and more common than those of inaction because it's typically easier to imagine undoing an action taken (and mentally returning to the status quo) than to imagine what would have resulted from an unchosen option (Kahneman, 1995). Other research paints a more complex picture, documenting a temporal shift in people's regrets over actions and inactions—namely, that regrets of action are more intense in the short term, but regrets of inaction gain prominence and stand out in the long run (Gilovich & Medvec, 1994, 1995; for an exception, see Morrison & Roese, 2011).

What you write in your introduction will be more effective and more enjoyable to read if it is framed in terms of ideas and findings rather than studies and scientists. It will also probably be shorter.

Punctuation

While you're learning to write, learn to punctuate. There are several punctuation marks that are the friend of any serious writer. These are: the em dash, the colon, the semi-colon, the comma, and—perhaps surprisingly—the period.

In *On Writing Well* (2016), William Zinsser says, "There's not much to be said about the period except that most writers don't reach it often enough." We all can learn from that.

Ending Strong

Just as you want to start strong, you should end strong. The best way to do that is to identify the take home message from your work. What has your study uncovered? What does the reader need to remember? Restating that idea at the end is a good way to finish.

One technique you can use to add a little flair to your ending is to return to whatever theme, idea, or technique you used to grab the reader's interest in the first place. If you posed a rhetorical question at the start of your paper, maybe you can answer that question at the end. If you observed something about behavior, perhaps you have something more to say about that behavior in light of your results. Tying your ending to the hook you used to grab readers attention is a nice way to round out your paper and send readers on their way with a satisfying flourish.

Recommended Books on Writing

Learning to write well is a process. The books below discuss many of the ideas above in more depth and offer a good starting point for students who are serious about learning to write non-fiction.

- *Write it Up* by Paul Silvia
- *Scientific Writing for Psychology* by Robert V. Kail
- *On Writing Well* by William Zinsser
- *The Sense of Style* by Steven Pinker
- *Writing to Learn* by William Zinsser

Preview

References

- Adachi, P. J. C., & Willoughby, T. (2011). The effect of violent video games on aggression: Is it more than just the violence? *Aggression and Violent Behavior, 16*(1), 55-62.
- Adelson, J. L., & McCoach, D. B. (2010). Measuring the mathematical attitudes of elementary students: The effects of a 4-point or 5-point Likert-type scale. *Educational and Psychological Measurement, 70*(5), 796-807. <https://doi.org/10.1177/0013164410366694>
- Alderotti, G., Rapallini, C., & Traverso, S. (2023). The Big Five personality traits and earnings: A meta-analysis. *Journal of Economic Psychology, 94*, 102570. <https://doi.org/10.1080/00221325.2023.2279143>
- Allik, J., & McCrae, R. R. (2004). Toward a geography of personality traits: Patterns of profiles across 36 cultures. *Journal of cross-cultural psychology, 35*(1), 13-28.
- Allen, E. J., Dechow, P. M., Pope, D. G., & Wu, G. (2015). Reference-dependent preferences: Evidence from marathon runners. *Management Science, 63*(6), 1657-1672. <https://doi.org/10.1287/mnsc.2015.2417>
- American Psychological Association. (2020). *Publication manual of the American Psychological Association 2020: The Official Guide to APA Style (7th ed.)*. American Psychological Association.
- Anderson, C. A., & Bushman, B. J. (2001). Effects of Violent Video Games on Aggressive Behavior, Aggressive Cognition, Aggressive Affect, Physiological Arousal, and Prosocial Behavior: A Meta-Analytic Review of the Scientific Literature. *Psychological Science, 12*(5), 353-359. <https://doi.org/10.1111/1467-9280.00366>
- Anderson, C. A., & Bushman, B. J. (2002). Human aggression. *Annual Review of Psychology, 53*, 27-51.
- Anderson, C. A., Allen, J. J., Plante, C., Quigley-McBride, A., Lovett, A., & Rokkum, J. N. (2019). The MTurkification of social and personality psychology. *Personality and Social Psychology Bulletin, 45*(6), 842-850. <https://doi.org/10.1177/0146167218798821>

- Anderson, E., & Durstine, J. L. (2019). Physical activity, exercise, and health: Quantifying the benefits. *American Journal of Lifestyle Medicine*, 13(5), 430-442.
- Arechar, A. A., Gächter, S., & Molleman, L. (2018). Conducting interactive experiments online. *Experimental Economics*, 21(1), 99–131. <https://doi.org/10.1007/s10683-017-9527-2>
- Arechar, A. A., Kraft-Todd, G. T., & Rand, D. G. (2017). Turking overtime: How participant characteristics and behavior vary over time and day on Amazon Mechanical Turk. *Journal of the Economic Science Association*, 3, 1–11. <https://doi.org/10.1007/s40881-017-0035-0>
- Arndt, A. D., Ford, J. B., Babin, B. J., & Luong, V. (2022). Collecting samples from online services: How to use screeners to improve data quality. *International Journal of Research in Marketing*, 39(1), 117-133.
- Ashton, M. C., & Lee, K. (2007). Empirical, theoretical and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review*, 11(2), 150–166.
- Austin, E. A., et al. (2024) - Reference for reliability of AI for qualitative data analysis
- Austin, E. A., et al. (2025) - Reference for once for AI conducting in-depth interviews with many participants simultaneously, and once in the summary for AI systems combining structured surveys, in-depth interviews, and large-sample statistical analysis
- Awad, E., Dsouza, S., Sharif, A., Rahwan, I., & Bonnefon, J. F. (2020). Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the National Academy of Sciences*, 117(5), 2332-2337.
- Bai, H. (2018). Evidence that a large amount of low quality responses on MTurk can be detected with repeated GPS coordinates. Retrieved from:
<https://www.maxhuibai.com/blog/evidence-that-responsesfrom-repeating-gps-are-random>
- Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., ... & Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of survey statistics and methodology*, 1(2), 90-143.

- Bastien, C. H., Vallières, A., & Morin, C. M. (2001). Validation of the Insomnia Severity Index as an outcome measure for insomnia research. *Sleep medicine*, 2(4), 297-307.
- Batson, C. D., Early, S., & Salvarani, G. (1997). Perspective taking: Imagining how another feels versus imagining how you would feel. *Personality and Social Psychology Bulletin*, 23(7), 751-758. <https://doi.org/10.1177/0146167297237008>
- Batson, C. D., Lishner, D. A., Cook, J., & Sawyer, S. (2005). Similarity and nurturance: Two possible sources of empathy for strangers. *Basic and Applied Social Psychology*, 27(1), 15–25. https://doi.org/10.1207/s15324834basp2701_2
- Benet-Martínez, V., & John, O. P. (1998). Los Cinco Grandes across cultures and ethnic groups: Multitrait-multimethod analyses of the Big Five in Spanish and English. *Journal of personality and social psychology*, 75(3), 729.
- Benjamin Jr, L. T., & Simpson, J. A. (2009). The power of the situation: The impact of Milgram's obedience studies on personality and social psychology. *American psychologist*, 64(1), 12.
- Benoit, K., Conway, D., Lauderdale, B. E., Ober, M., & Mikhaylov, S. (2016). Crowd-sourced text analysis: Reproducible and large production of political data. *American Political Science Review*, 110(2), 283-295.
- Bhattacharjee, Y. (2013). The mind of a con man. *The New York Times*. <https://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html>
- Bogg, T., & Roberts, B. W. (2013). The case for conscientiousness: Evidence and implications for a personality trait marker of health and longevity. *Annals of Behavioral Medicine*, 45(3), 278-288.
- Bogle, K. A. (2008). *Hooking up: Sex, dating, and relationships on campus*. New York University Press.
- Bohannon J. Mechanical Turk upends social sciences. *Science*. 2016; 352 (6291): 1263–1264. pmid:27284175

- Bolger, N., & Laurenceau, J. P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. Guilford press.
- Boynton, M. H., & Richman, L. S. (2014). An online daily diary study of alcohol use using Amazon's Mechanical Turk: Daily diary study of alcohol using MTurk. *Drug and Alcohol Review*, 33(4), 456–461. <https://doi.org/10.1111/dar.12163>
- Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3, 296–322
- Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data?. *Perspectives on Psychological Science*, 6(1), 3-5. <https://doi.org/10.1177/1745691610393980>
- Buhrmester, M. D., Talaifar, S., & Gosling, S. D. (2018). An evaluation of Amazon's Mechanical Turk, its rapid rise, and its effective use. *Perspectives on psychological science*, 13(2), 149-154.
- Campbell, D. S., & Reiman, A.-K. (2022). Has social psychology lost touch with reality? Exploring public perceptions of the realism and consequentiality of social psychological research. *Journal of Experimental Social Psychology*, 98, 104255. <https://doi.org/10.1016/j.jesp.2021.104255>
- Casey, L. S., Chandler, J., Irvine, A. S., Proctor, A., & Strolovitch, D. Z. (2017). Intertemporal differences among MTurk workers: Time-based sample variations and implications for online data collection. *SAGE Open*, 7,1–15. <https://doi.org/10.1177/2158244017712774>
- Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. (1970). *Handbook for the Sixteen Personality Factor Questionnaire (16PF)*. Champaign, IL: IPAT
- Chan, J. C. (1991). Response-Order Effects in Likert-Type Scales. *Educational and Psychological Measurement*, 51(3), 531-540. <https://doi.org/10.1177/0013164491513002>
- Chandler, J., Litman, L., & Robinson, J. (2021). Maximum retention rates in longitudinal studies on Mechanical Turk. Unpublished manuscript.

- Chandler, J. J., & Paolacci, G. (2017). Lie for a dime: When most prescreening responses are honest but most study participants are impostors. *Social Psychological and Personality Science*, 8(5), 500-508. <https://doi.org/10.1177/1948550617698203>
- Chandler, J., Rosenzweig, C., Moss, A. J., Robinson, J., & Litman, L. (2019). Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior research methods*, 51, 2022-2038.
- Chandler, J., Sisso, I., & Shapiro, D. (2020). Participant carelessness and fraud: Consequences for clinical research and potential solutions. *Journal of Abnormal Psychology*, 129(1), 49. <https://doi.org/10.1037/abn0000479>
- Cherry, G., Rocke, J., Chu, M., Liu, J., Lechner, M., Lund, V. J., & Kumar, B. N. (2020). Loss of smell and taste: a new marker of COVID-19? Tracking reduced sense of smell during the coronavirus pandemic using search trends. *Expert Review of Anti-Infective Therapy*, 18(11), 1165-1170.
- Chmielewski, M., & Kucker, S. C. (2020). An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, 11(4), 464-473.
- Chopra, F., & Haaland, I. (2023). Combining qualitative interviews with AI. <http://dx.doi.org/10.1177/10939839231184583756>
- Chyung, S. Y., Kennerly, M., & Campbell, I. (2018). Evidence-based survey design: The use of ascending or descending order of Likert-type response options. *Performance Improvement*, 57(9), 9-16. <https://doi.org/10.1002/pfi.21800>
- Cimpian, J. R., & Timmer, J. D. (2020). Mischievous responders and sexual minority youth survey data: A brief history, recent methodological advances, and implications for research and practice. *Archives of Sexual Behavior*, 49(4), 1097-1102. <https://doi.org/10.1007/s10508-020-01661-7>
- Clifford, S., Jewell, R. M., & Waggoner, P. D. (2015). Are samples drawn from Mechanical Turk valid for research on political ideology?. *Research & Politics*, 2(4), 2053168015622072.

- Clifton, J. D., & Kerry, N. (2023). Belief in a dangerous world does not explain substantial variance in political attitudes, but other world beliefs do. *Social Psychological and Personality Science*, 14(5), 515-525.
- Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science*, 1(3), 98-101. <https://doi.org/10.1111/1467-8721.ep10768783>
- Cohn, N. (2024, September 27). *The problem with a crowd of new online polls*. The New York Times. <https://www.nytimes.com/2024/09/27/upshot/online-polls-election.html>
- Converse, B. A., & Epley, N. (2007). The anchoring effect of God: Religious priming and attitude change. Unpublished manuscript, University of Chicago.
- Coppock, A., Leeper, T. J., & Mullinix, K. J. (2018). Generalizability of heterogeneous treatment effect estimates across samples. *Proceedings of the National Academy of Sciences*, 115(49), 12441-12446.
- Costa, P. T., & McCrae, R. R. (2008). The revised neo personality inventory (neo-pi-r). *The SAGE handbook of personality theory and assessment*, 2(2), 179-198.
- Costello, T. H., Pennycook, G., & Rand, D. G. (2025, March 7). Just the facts: How dialogues with AI reduce conspiracy beliefs. https://doi.org/10.31234/osf.io/h7n8u_v2
- Couper M. P., Traugott M., Lamias M. (2001). Web survey design and administration. *Public Opinion Quarterly*, 65, 230–253. <https://doi.org/10.1086/322199>
- Couper, M. P., Tourangeau, R., Conrad, F. G., & Zhang, C. (2013). The design of grids in web surveys. *Social science computer review*, 31(3), 322-345.
- Cromer, L. D., & Newman, E. (2011). Research ethics in victimization studies: Widening the lens. *Violence against Women*, 17(12), 1489-1508. <https://doi.org/10.1177/1077801211436365>
- Curry, L., & Nunez-Smith, M. (2015). *Mixed methods in health sciences research: A practical primer* (Vol. 1). Sage publications.

- Davis, M. H., Conklin, L., Smith, A., & Luce, C. (1996). Effect of perspective taking on the cognitive representation of persons: a merging of self and other. *Journal of personality and social psychology*, 70(4), 713.
- Day, J. C., & Newburger, E. C. (2002). The Big Payoff: Educational Attainment and Synthetic Estimates of Work-Life Earnings. Special Studies, Current Population Reports. U.S. Census Bureau.
- Debell, M., Wilson, C., Jackman, S., & Figueroa, L. (2021). Optimal response formats for online surveys: Branch, grid, or single item?. *Journal of Survey Statistics and Methodology*, 9(1), 1-24. <https://doi.org/10.1093/jssam/smz039>
- Dennis, S. A., Goodson, B. M., & Pearson, C. A. (2020). Online worker fraud and evolving threats to the integrity of MTurk data: A discussion of virtual private servers and the limitations of IP-based screening procedures. *Behavioral Research in Accounting*, 32(1), 119-134. <https://doi.org/10.2308/bria-18-044>
- DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2013). Best practice recommendations for data screening. *Journal of Organizational Behavior*, 36(2), 171-181.
- Donders, A. R. T., Van Der Heijden, C. M., Stijnen, T., & Moons, K. G. (2006). A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, 59(10), 1087-1091. <https://doi.org/10.1016/j.jclinepi.2006.01.014>
- Donnellan, M. B., & Rakhmani, A. (2023). How does the number of response options impact the psychometric properties of the Rosenberg Self-Esteem Scale?. *Assessment*, 30(6), 1737-1749.
- Douglas, B. D., McGorray, E. L., & Ewell, P. J. (2021). Some researchers wear yellow pants, but even fewer participants read consent forms: Exploring and improving consent form reading in human subjects research. *Psychological methods*, 26(1), 61.
- Dubow, E. F., Boxer, P., & Huesmann, L. R. (2009). Long-term effects of parents' education on children's educational and occupational success: Mediation by family interactions, child

- aggression, and teenage aspirations. *Merrill-Palmer quarterly* (Wayne State University Press), 55(3), 224.
- Epley, N., & Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological Science*, 12(5), 391-396. <https://doi.org/10.1111/1467-9280.00372>
- Fan, X., Miller, B. C., Park, K. E., Winward, B. W., Christensen, M., Grotevant, H. D., & Tai, R. H. (2006). An exploratory study about inaccuracy and invalidity in adolescent self-report surveys. *Field methods*, 18(3), 223-244.
- Fiske, D. W. (1949). Consistency of the factorial structures of personality ratings from different sources. *Journal of Abnormal and Social Psychology*, 44 (1), 329-344. <https://doi.org/10.1037/h0057198>
- Fordsham, N., Moss, A. J., Krumholtz, S., Roggiana, T., & Robinson, J., & Litman, L. (2019). Variation among Mechanical Turk workers across time of day presents an opportunity and a challenge for research. *PsyArXiv*. <https://doi.org/10.31234/osf.io/p8bns>
- Fowler, C., Jiao, J., & Pitts, M. (2023). Frustration and ennui among Amazon MTurk workers. *Behavior Research Methods*, 55(6), 3009-3025. <https://doi.org/10.1177/1063426921119532>
- Friedman, L. W., & Friedman, H. H. (1994). A comparison of vertical and horizontal rating scales. *The Mid-Atlantic Journal of Business*, 30(1), 107. <https://doi.org/10.1177/0894439319834296>
- Galesic, M., Tourangeau, R., Couper, M. P., & Conrad, F. G. (2007, March). Using change to improve navigation in grid questions. In *German Online Research Conference*, Leipzig, Germany.
- Galinsky, A. D., & Moskowitz, G. B. (2000). Perspective-taking: Decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of Personality and Social Psychology*, 78(4), 708-724.

- Gallo, J., & Gran-Ruaz, S. (2021, October 1). Racial Trauma Scale: Creative interview strategies employed in the development of a new clinical tool for measuring race-based stress and trauma. *Innovations in Online Research, Virtual*.
https://www.youtube.com/watch?v=hKCK_dWfUfl&t=1568s
- Garbarski, D., Schaeffer, N.C. & Dykema, J. (2019). The Effects of features of survey measurement on self-rated health: Response option order and scale orientation. *Applied Research Quality Life*, 14, 545–560. <https://doi.org/10.1007/s11482-018-9628-x>
- Garbinsky, E. N., Gladstone, J. J., Nikolova, H., & Olson, J. G. (2020). Love, lies, and money: Financial infidelity in romantic relationships. *Journal of Consumer Research*, 47(1), 1-24.
<https://doi.org/10.1093/jcr/ucz052>
- George, D., & Mallery, P. (2003). *SPSS for Windows Step by Step: A Simple Guide and Reference* (4th ed.). Boston: Allyn & Bacon.
- Gharpure, R., Hunter, C. M., Schnall, A. H., et al. (2020). Knowledge and practices regarding safe household cleaning and disinfection for COVID-19 prevention — United States, May 2020. *MMWR. Morbidity and Mortality Weekly Report*, 69(23), 705–709.
<https://doi.org/10.15585/mmwr.mm6923e2>
- Gilbert, S. J. (1981). Another look at the Milgram obedience studies: The role of the graded series of shocks. *Personality and Social Psychology Bulletin*, 7(4), 690-695.
- Gleaves, D. H., Rucklidge, J., & Follette, V. M. (2007). What are we teaching our students by not asking about abuse? *American Psychologist*, 62(4), 326–327. <https://doi.org/10.1037/0003-066X62.4.326>
- Goldberg, L. R. (1981). Language and individual differences: The search for universals in personality lexicons. *Review of personality and social psychology*, 2(1), 141-165.
- Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, 37 (6), 504-528.
[https://doi.org/10.1016/S0092-6566\(03\)00046-1](https://doi.org/10.1016/S0092-6566(03)00046-1)

- Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology*, 47, 55-130.
- Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Spassena, K., & Ditto, P. H. (2008). Moral foundations questionnaire. *Journal of Personality and Social Psychology*.
<https://doi.org/10.1037/t05651-000>
- Grandmont, J., Graff, B., Goetzinger, L., & Dorbecker, K. (2010, May). Grappling with grids: How does question format affect data quality and respondent engagement. In Annual meeting of the American Association for Public Opinion Research. Retrieved from <http://www.amstat.org/sections/SRMS/proceedings/y2010/File/400116.pdf>.
- Gray, K., Schein, C., & Ward, A. F. (2014). The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology: General*, 143(4), 1600. <http://dx.doi.org/10.1037/xap0036129>
- Greitemeyer, T., & Mügge, D. O. (2014). Video games do affect social outcomes: A meta-analytic review of the effects of violent and prosocial video game play. *Personality and Social Psychology Bulletin*, 40(5), 578-589.
- Griffin, M. G., Resick, P., Wainrop, A. E., & Mechanic, M. B. (2003). Participation in trauma research: Is there evidence of harm? *Journal of Traumatic Stress*, 16, 221-227.
- Griggs, R. A. (2017). Milgram's obedience study: A contentious classic reinterpreted. *Teaching of Psychology*, 44(1), 32-37.
- Gummer, T., Roßmann, J., & Silber, H. (2021). Using instructed response items as attention checks in web surveys: Properties and implementation. *Sociological Methods & Research*, 50(1), 238–264. <https://doi.org/10.1177/0049124118769083>
- Hall, M. P., Lewis, Jr., N. A., Chandler, J., & Litman, L. (2020). Conducting longitudinal research on Amazon Mechanical Turk. In L. Litman and J. Robinson (Eds.) *Conducting Online Research on Amazon Mechanical Turk and Beyond* (198-216). Sage Academic Publishing. Thousand Oaks, CA.

- Harari, G. M., Stachl, C., Müller, S. R., & Gosling, S. D. (2021). Mobile sensing for studying personality dynamics in daily life. In *The Handbook of Personality Dynamics and Processes* (pp. 763-790). Academic Press.
- Hartman, R., Robinson, J., Moss, A., & Litman, L. (2023, March 09). How Do Black Americans Really Feel about the Phrase "It's ok to be White"? CloudResearch Blog.
<https://www.cloudresearch.com/resources/blog/its-ok-to-be-white-rasmussen-poll/>
- Hartman, R., Moss, A. J., Rabinowitz, I., Bahn, N., Rosenzweig, C., Robinson, J., & Litman, L. (2023). Do you know the Woolly Bully? Testing era-based knowledge to verify participant age online. *Behavior research methods*, 55(7), 3313-3325.
- Hartman, R., Moss, A. J., Jaffe, S. N., Rosenzweig, C., Litman, L., & Robinson, J. (2023). Introducing Connect by CloudResearch: Advancing online participant recruitment in the digital age. *PsyArXiv*.
- Hauser, D. J., Ellsworth, P. C., & Gonzalez, R. (2016). Are manipulation checks necessary?. *Frontiers in psychology*, 9, 998.
- Hauser, D. J., & Schwarz, N. (2015). It's a trap! Instructional manipulation checks prompt systematic thinking on "tricky" tasks. *Science Open*, 5(2), 2158244015584617.
- Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, 48, 400-407. <https://doi.org/10.3758/s13428-015-0578-z>. PMID: 25761395.
- Hauser, D. J., Moss, A. J., Rosenzweig, C., Jaffe, S. N., Robinson, J., & Litman, L. (2023). Evaluating CloudResearch's Approved Group as a solution for problematic data quality on MTurk. *Behavior Research Methods*, 55(8), 3953-3964.
- Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Publications.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world?. *Behavioral and brain sciences*, 33(2-3), 61-83.

- Holliday, D. E., Iyengar, S., Lelkes, Y., & Westwood, S. J. (2024). Uncommon and nonpartisan: Antidemocratic attitudes in the American public. *Proceedings of the National Academy of Sciences*, 121(13), e2313013121.
- Hu, J. (2020). Horizontal or vertical? The effects of visual orientation of categorical response options on survey responses in web surveys. *Social Science Computer Review*, 38(6), 779-792.
- Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, 27(1), 99–114. <https://doi.org/10.1007/s10869-011-9231-8>
- Huang, J. L., Bowling, N. A., Liu, M., & Li, Y. (2015). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology*, 30, 299-311.
- Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, 100(3), 828–845. <https://doi.org/10.1037/a0038510>
- Hunt, M. G., Marx, R., Lipson, C., & Bowling, J. (2018). No more FOMO: Limiting social media decreases loneliness and depression. *Journal of Social and Clinical Psychology*, 37(10), 751-768. <http://doi.org/10.1521/jscp.2018.37.10.751>
- Hyde, A. L., Conroy, D. E., Pincus, A. L., & Ram, N. (2011). Unpacking the feel-good effect of free-time physical activity: Between-and within-person associations with pleasant-activated feeling states. *Journal of Sport and Exercise Psychology*, 33(6), 884-902.
- Iglesias, C. P., Birks, Y. F., & Torgerson, D. J. (2001). Improving the measurement of quality of life in older people: The York SF-12. *Quarterly Journal of Medicine*, 94, 695–698. Retrieved from <http://doi.org/10.1093/qjmed/94.12.695>
- Jackson, D. N. (1976). The appraisal of personal reliability. Paper presented at the meetings of the Society of Multivariate Experimental Psychology, University Park, PA.

- Jaffe, S. N., Moss, A. J., Rosenzweig, C., Gautam, R., Robinson, J., & Litman, L. (2025). The bots ruining social science aren't bots at all. Manuscript under review.
- John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 102-138). Guilford Press.
- John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of Personality: Theory and Research* (3rd ed., pp. 114–158). Guilford Press.
- Johns, R. (2005) One size doesn't fit all: Selecting response scales for attitude items. *Journal of Elections, Public Opinion and Parties*, 15(2), 237–264.
<https://doi.org/10.1080/136898805001788>
- Judge, T. A., Rodell, J. B., Klinger, R. L., Simon, L. S., & Crawford, E. R. (2013). Hierarchical representations of the five-factor model of personality in predicting job performance: integrating three organizing frameworks with two theoretical perspectives. *Journal of applied psychology*, 98(6), 875–896.
- Judge, T. A., Thoresen, C. J., Bono, J. E., & Patton, G. K. (2001). The job satisfaction-job performance relationship: A qualitative and quantitative review. *Psychological Bulletin*, 127(3), 376-407. <https://doi.org/10.1037/0033-2909.127.3.376>
- Kaczmirek, L. (2008). Human-survey interaction: Usability and nonresponse in online surveys. Retrieved from <https://ub-madoc.bib.uni-mannheim.de/2150>
- Kaczmirek, L. (2011). Attention and usability in Internet surveys: Effects of visual feedback in grid questions. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies*. (pp. 191–214). Routledge/Taylor & Francis Group.
- Kail, R. V. (2015). *Scientific writing for psychology: Lessons in clarity and style*. Sage Publications, Inc.

- Kalin, N. H. (2020). The critical relationship between anxiety and depression. *American Journal of Psychiatry*, 177 (5), 365-367.
- Kane, J. V., Velez, Y. R., & Barabas, J. (2023). Analyze the attentive and bypass bias: Mock vignette checks in survey experiments. *Political Science Research and Methods*, 11(2), 293-310.
- Katz, J., Sanger-Katz, M., & Quealy, K. (2020, July 17). A detailed map of who is wearing masks in the U.S. *The New York Times*. Retrieved from <https://www.nytimes.com/interactive/2020/07/17/upshot/coronavirus-face-mask-map.html>
- Keles, B., McCrae, N., & Grealish, A. (2020). A systematic review of the influence of social media on depression, anxiety and psychological distress in adolescents. *International journal of adolescence and youth*, 25(1), 79-93.
- Keller, L. & Kabengele, M. C. (2022, May 16). Research on the regulation of face-touching. *CloudResearch*. <https://www.cloudresearch.com/resources/blog/self-regulation-of-face-touching-remote-observation/>
- Kennedy, C. (2018, August 6). *What are nonprobability surveys?* Pew Research Center. <https://www.pewresearch.org/short-reads/2018/08/06/what-are-nonprobability-surveys/>
- Kennedy, C., Hatley, T., La, A., Mercer, A., Keeter, A., Fern, J., & Asare-Marfo, D. (2020, February 18). *Assessing the risks to online polls from bogus respondents*. Pew Research Center. https://www.pewresearch.org/methods/2020/02/18/assessing-the-risks-to-online-polls-from-bogus-respondents/?utm_source=chatgpt.com#fn-101287-1
- Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P. D., Jewell, R., & Winter, N. J. (2020). The shape of and solutions to the MTurk quality crisis. *Political Science Research and Methods*, 8(4), 614-629.
- Kerry, N., Prokosch, M. L., & Murray, D. R. (2023). The Holy Father (and Mother)? Multiple tests of the hypothesis that parenthood and parental care motivation lead to greater

- religiosity. *Personality and Social Psychology Bulletin*, 49(5), 709-726.
<https://doi.org/10.1177/01461672221076919>
- Kingsbury, K. (2025, April 17). To understand global migration, you have to see it first. *The New York Times*. <https://www.nytimes.com/interactive/2025/04/17/opinion/global-migration-facebook-data.html>
- Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. *Proceeding of the Twenty-Sixth Annual CHI Conference on Human Factors in Computing Systems - CHI '08*, 453. <https://doi.org/10.1145/1357054.1357127>
- Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., ... & Sowden, W. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4), 443-490. <https://doi.org/10.1177/2515245918810225>
- Kohlberg, L. (1984) *The psychology of moral development: Essays on moral development*. Harper & Row, San Francisco.
- Kraut, R., Olson, J., Banaji, M., Bruckman, J., Cohen, J., & Couper, M. (2004). Psychological research online: report of Panel on Scientific Affairs' Advisory Group on the Conduct of Research on the Internet. *American psychologist*, 59(2), 105.
- Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9), 606-613.
- Kulas, J. T., & Stachowski, A. A. (2009). Middle category endorsement in odd-numbered Likert response scales: Associated item characteristics, cognitive demands, and preferred meanings. *Journal of Research in Personality*, 43(3), 489-493.
<https://doi.org/10.1016/j.jrp.2008.12.005>
- Kung, F. Y., Kwok, N., & Brown, D. J. (2017). Are attention check questions a threat to scale validity?. *Applied Psychology*, 67(2), 264-283. <https://doi.org/10.1111/apps.12108>
- Ledgerwood, A., da Silva Frost, A., Kadirvel, S., Maitner, A. T., Wang, Y. A., & Maddox, K. B. (in press). Methods for advancing an open, replicable, and inclusive science of social

- cognition. Chapter to appear in K. Hugenberg, K. Johnson, & D. E. Carlston (Eds), Oxford Handbook of Social Cognition.
- Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate behavioral research*, 39(2), 329-358.
- Levinson, D. J. (1986). *The Seasons of a Man's Life: The Groundbreaking 10-Year Study That Was the Basis for Passages!*. Ballantine Books.
- Lewis, M. (2004). *Moneyball: The art of winning an unfair game*. WW Norton.
- Liebow, E. (1993). *Tell them who I am: The lives of homeless women*. Penguin Books.
- Litman, L., & Robinson, J. (2020). *Conducting online research on Amazon Mechanical Turk and beyond*. Sage Publications.
- Litman, L., & Moss, A. J. (2020, July 24). What is it like to participate in online research on Amazon Mechanical Turk? CloudResearch Blog. <https://www.cloudresearch.com/resources/blog/page/12/>
- Litman, L., & Robinson, J. (2020). Conducting ethical online research: A data-driven approach. In L. Litman and J. Robinson (Eds.) *Conducting Online Research on Amazon Mechanical Turk and Beyond* (234-263). Sage Academic Publishing. Thousand Oaks: CA.
- Litman, L., Robinson, J., & Rosenzweig, C. (2015). The relationship between motivation, monetary compensation, and data quality among US-and India-based workers on Mechanical Turk. *Behavior research methods*, 47(2), 519-528.
- Litman, L., Moss, A. J., Rosenzweig, C., Jaffe, S. N., Gautam, R., & Robinson, J. (2021). Bots or inattentive humans? Identifying sources of low-quality data in online platforms. PsyArXiv. <https://doi.org/10.31234/osf.io/wr8ds>
- Litman, L., Rosen, Z., Hartman, R., Rosenzweig, C., Weinberger-Litman, S. L., Moss, A. J., & Robinson, J. (2023). Did people really drink bleach to prevent COVID-19? A guide for protecting survey data against problematic respondents. *Plos one*, 18(7), e0287837.

- Liu, M., & Cernat, A. (2018). Item-by-item versus matrix questions: A web survey experiment. *Social Science Computer Review*, 36(6), 690-706.
- Liu, M., & Wronski, L. (2018). Examining completion rates in web surveys via over 25,000 real-world surveys. *Social Science Computer Review*, 36(1), 116-124.
- Löwe, B., Decker, O., Müller, S., Brähler, E., Schellberg, D., Herzog, W., & Herzberg, P. Y. (2008). Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. *Medical Care*, 46(3), 266-274. DOI: 10.1097/MLR.0b013e318160d093
- Luyster, F. S., Strollo Jr, P. J., Zee, P. C., & Walsh, J. K. (2012). Sleep: A health imperative. *Sleep*, 35(6), 727-734. <https://doi.org/10.5665/sleep.1846>
- Malouff, J. M., Schutte, N. S., & Thorsteinsson, E. B. (2014). Trait emotional intelligence and romantic relationship satisfaction: A meta-analysis. *The American Journal of Family Therapy*, 42 (1), 53-66. <https://doi.org/10.1002/ajft.2012.748549>
- Mark, G., Gudith, D., & Klocke, U. (2008). The cost of interrupted work: More speed and stress. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 107-110.
- Mark, G., Iqbal, S., Czornyj, M., & Johns, P. (2016). Bored Mondays and focused afternoons: The rhythm of attention and online activity in the workplace. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 3539-3548.
- Markey, P. M., & Markey, C. N. (2010). Vulnerability to violent video games: A review and integration of personality research. *Review of General Psychology*, 14(2), 82-91.
- McCrae, R. R., & Costa Jr, P. T. (1997). Personality trait structure as a human universal. *American psychologist*, 52(5), 509.
- McCrae, R. R., Terracciano, A., & 78 Members of the Personality Profiles of Cultures Project. (2005). Universal Features of Personality Traits From the Observer's Perspective: Data From 50 Cultures. *Journal of Personality and Social Psychology*, 88(3), 547-561. <https://doi.org/10.1037/0022-3514.88.3.547>

- Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437–455. <https://doi.org/10.1037/a0028085>
- Medina, E. (2023, February 26). Newspapers Drop 'Dilbert' After Creator's Rant About Black 'Hate Group'. *The New York Times*. <https://www.nytimes.com/2023/02/26/us/dilbert-newspapers-racism.html>
- Mercer, A. Kenedy, C., Keeter, S. (2024, March 5). Online opt-in polls can produce misleading results, especially for young people and Hispanic adults. Pew Research Center. [blog]. <https://www.pewresearch.org/short-reads/2024/03/05/online-opt-in-polls-can-produce-misleading-results-especially-for-young-people-and-hispanic-adults/>
- Milgram, S. (1963). Behavioral study of obedience. *The Journal of Abnormal and Social Psychology, 67*, 371-378. <http://dx.doi.org/10.1037/h0040527>
- Miller, B. C., Fan, X., Christensen, M., Grotevant, H. D., & Van Dulmen, M. (2000). Comparisons of adopted and nonadopted adolescents in a large, nationally representative sample. *Child development, 71*(5), 1458-1473.
- Molloy, G. J., O'Carroll, R. E., & Ferguson, J. (2014). Conscientiousness and medication adherence: A meta-analysis. *Annals of Behavioral Medicine, 47*(1), 92-101.
- Morehouse, K., Kurdi, B., & Hošek, R. A. (2023). Responsible data sharing: Identifying and remedying possible re-identification of human participants. <https://doi.org/10.3122/osf.io/5m3cx>
- Moss, A., & Litman, L. (2018). After the bot scare: Understanding what's been happening with data collection on MTurk and how to stop it. Retrieved February, 4, 2019.
- Moss, A. J. (February, 2024). What's "fair" compensation in online research? The Society for Personality and Social Psychology Annual Convention, San Diego, CA.
- Moss, A. J. (2022, March 04). How CloudResearch and IARPA completed the largest longitudinal online research project ever. CloudResearch. <https://www.cloudresearch.com/resources/blog/the-largest-longitudinal-online-research-project/>

- Moss, A. J., Rosenzweig, C., Robinson, J., Jaffe, S. N., & Litman, L. (2023). Is it ethical to use Mechanical Turk for behavioral research? Relevant data from a representative survey of MTurk participants and wages. *Behavior Research Methods*, 55(8), 4048-4067.
- Moss, A. J., Hauser, D., Rosenzweig, C., Jaffe, S. N., Robinson, J., & Litman, L. (2023). Using market-research panels for behavioral science: An overview and tutorial. *Advances in Methods and Practices in Psychological Science*, 6(2), 25152459221140388.
- Moss, A. J., Hauser, D. J., Rosenzweig, C., Jaffe, S., Robinson, J., & Litman, L. (2024). Mechanical Turk: A Versatile Tool in the Behavioral Scientist's Toolkit. In A. L. Nichols & J. Edlund (Eds.), *The Cambridge Handbook of Research Methods and Statistics for the Social and Behavioral Sciences*. <https://doi.org/10.1017/9781009010054>
- Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, J. (2015). The Generalizability of Survey Experiments. *Journal of Experimental Political Science*, 2(2), 109–138. doi:10.1017/XPS.2015.19
- Nadler, J. T., Weston, R., & Voyles, E. C. (2015). Stuck in the middle: the use and interpretation of mid-points in items on questionnaires. *Journal of General Psychology*, 142(2), 71-89. <https://doi.org/10.1080/00221309.2014.994590>
- National Institute of Mental Health. (2023, March). Mental illness. Retrieved from: <https://www.nimh.nih.gov/health/statistics/mental-illness>
- National Survey on Drug Use and Health (n.d.) National Survey on Drug Use and Health. (n.d.). About the National Survey on Drug Use and Health. https://nsduhweb.rti.org/respweb/about_nsduh.html
- Neel, R., & Lassetter, B. (2019). The stigma of perceived irrelevance: An affordance-management theory of interpersonal invisibility. *Psychological Review*, 126(5), 634–659. <https://doi.org/10.1037/rev0000143>
- Newman, E., & Kaloupek, D. G. (2004). The risks and benefits of participating in trauma-focused research studies. *Journal of Traumatic Stress: Official Publication of The International Society for Traumatic Stress Studies*, 17(5), 383-394.

- Newman, E., & Kaloupek, D. (2009). Overview of research addressing ethical dimensions of participation in traumatic stress studies: Autonomy and beneficence. *Journal of Traumatic Stress: Official Publication of The International Society for Traumatic Stress Studies*, 22(6), 595-602.
- Newton, J. T., & Buck, D. J. (2000). Anxiety and pain measures in dentistry: a guide to their quality and application. *The Journal of the American Dental Association*, 131(10), 1449-1457. DOI: 10.14219/jada.archive.2000.0056
- Nomura, T., Suzuki, T., Kanda, T., & Kato, K. (2006). Measurement of anxiety toward robots. In *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 372-377). IEEE.
- Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Collier, F. S., Dreber, A., ... & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, 73, 719-748. <https://doi.org/10.1146/annurev-psych-020821-114157>
- O'Grady, C. (2024, May 9). Psychology study participants recruited online may provide nonsensical answers. *Science*. <https://www.science.org/content/article/psychology-study-participants-recruited-online-may-provide-nonsensical-answers>
- Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4), 867-872. <https://doi.org/10.1016/j.jesp.2009.03.009>
- Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5(5), 411-419. <https://doi.org/10.1017/S1930297500002205>
- Pape, H., & Storvoll, E. E. (2006). Teenagers'"use" of Non-Existent Drugs: A study of false positives. *Nordic Studies on Alcohol and Drugs*, 23(1_suppl), 31-46.
- Park, Y., MacDonald, G., Impett, E. A., & Neel, R. (2023). What social lives do single people want? A person-centered approach to identifying profiles of social motives among

- singles. *Journal of Personality and Social Psychology*, 125(1), 219.
<https://doi.org/10.1037/pspp0000455>
- Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, 46, 1023-1031.
<https://doi.org/10.3758/s13428-013-0434-y>
- Peer, E., Rothschild, D., & Gordon, A. (2023, March 22). Platform over procedure: Online platforms that pre-vet participants yield higher data quality without sacrificing diversity.
<https://doi.org/10.31234/osf.io/buzwn>
- Pereira-Doel, P., Daly, J. E. M., & Walker, I. (2024, March 7). Beyond the water flow rate: Water pressure and smart timers impact shower efficiency
<https://doi.org/10.31219/osf.io/hxaey>
- Peterson, R. A., & Merunka, D. R. (2014). Convenience sample of college students and research reproducibility. *Journal of Business Research*, 67(10), 1035-1041.
- Permut, S., Fisher, M., & Oppenheimer, D. M. (2010). TaskMaster: A tool for determining when subjects are on task. *Advances in Methods and Practices in Psychological Science*, 2(2), 188–196. <https://doi.org/10.1177/1515745919838479>
- Petzel, T. P., Johnson, J. E., & McKelip, J. (1973). Response bias in drug surveys. *Journal of Consulting and Clinical Psychology*, 40(3), 437–439. <https://doi.org/10.1037/h0034439>
- Phang, J., Lampe, M., Ahmar, L., Agarwal, S., Fang, C. M., Liu, A. R., ... & Maes, P. (2025). Investigating Affective Use and Emotional Well-being on ChatGPT. arXiv preprint arXiv:2504.03888.
- Pinker, S. (2015). *The sense of style: The thinking person's guide to writing in the 21st century* (No. 10). Penguin Books.
- Price, D. D., Finnis, D. G., & Benedetti, F. (2008). A comprehensive review of the placebo effect: Recent advances and current thought. *Annual Review of Psychology*, 59, 565-590.
<https://doi.org/10.1146/annurev.psych.59.113006.095941>

- Primack, B. A., Shensa, A., Sidani, J. E., Whaite, E. O., Yi Lin, L., Rosen, D., ... & Miller, E. (2017). Social media use and perceived social isolation among young adults in the US. *American Journal of Preventive Medicine*, 53(1), 1-8. <https://doi.org/10.1016/j.amepre.2017.01.010>
- Prins, A., Bovin, M. J., Smolenski, D. J., Marx, B. P., Kimerling, R., Jenkins-Guarnieri, M. A., ... & Tiet, Q. Q. (2016). The primary care PTSD screen for DSM-5 (PC-PTSD-5): development and evaluation within a veteran primary care sample. *Journal of general internal medicine*, 31(10), 1206-1211.
- Przybylski, A. K., & Weinstein, N. (2019). Violent video game engagement is not associated with adolescents' aggressive behaviour: evidence from a registered report. *Royal Society Open Science*, 6(2), 171474.
- Rakhshani, A., Donnellan, M. B., Roberts, B. W., & Lucas, R. E. (2023). Brief Report: Does the Number of Response Options Matter for the BFI-2? Conceptual Replication and Extension. *Assessment*. <https://doi.org/10.1177/10731911231190098>
- Rasmussen Reports. (2023, February 22). It's Okay to be White; Americans Overwhelmingly Agree with Hate "Speech" [Video]. YouTube. <https://www.youtube.com/watch?v=CvCJ8nV-myE>
- Reavey, B., Bruggeman, P., Rosenzweig, C., & Litman, L. (2024). Sentry In-Survey: A tool for preventing survey fraud. [Manuscript under review].
- Revilla, M., Toninelli, D., & Ochoa, C. (2017). An experiment comparing grids and item-by-item formats in web surveys completed through PCs and smartphones. *Telematics and Informatics*, 34(1), 30-42. <https://doi.org/10.1016/j.tele.2016.04.002>
- Ristow, T., & Hernandez, I. (2024). VOIS: A framework for recording Voice Over Internet Surveys. *Behavior Research Methods*, 56(1), 447-467. <https://doi.org/10.3758/s13428-022-02045-6>
- Rivera, E. D., Wilkowski, B. M., Moss, A. J., Rosenzweig, C., & Litman, L. (2022). Assessing the efficacy of a participant-vetting procedure to improve data-quality on Amazon's

- Mechanical Turk. *Methodology*, 18(2), 126-143.
<https://doi.org/10.1177/01461672231185484>
- Robinson, J. P., & Espelage, D. L. (2011). Inequities in Educational and Psychological Outcomes Between LGBTQ and Straight Students in Middle and High School. *Educational Researcher*, 40(7), 315-330. <https://doi.org/10.3102/0013189X11422112>
- Robinson, J. P., & Espelage, D. L. (2012). Bullying explains only part of LGBTQ–heterosexual risk disparities: Implications for policy and practice. *Educational Researcher*, 41(8), 309-319.
- Robinson, J. P., & Espelage, D. L. (2013). Peer victimization and sexual risk differences between lesbian, gay, bisexual, transgender, or questioning and non-transgender heterosexual youths in grades 7–12. *American journal of public health*, 103(10), 1810-1819.
- Robinson-Cimpian, J. P. (2014). Inaccurate estimation of disparities due to mischievous responders: Several suggestions to assess comprehension. *Educational Researcher*, 43(4), 171-185.
- Robinson, J., Rosenzweig, C., Moss, A. J., & Litman, L. (2019). Tapped out or barely tapped? Recommendations for how to harness the vast and largely unused potential of the Mechanical Turk participant pool. *PLoS ONE*, 14(12): e0226394.
<https://doi.org/10.1371/journal.pone.0226394>
- Rojanasakul, M., Flavelle, K., Migliozi, B., & Murray, E. (2023, August 28). America is using up its groundwater like there's no tomorrow. *The New York Times*.
<https://www.nytimes.com/interactive/2023/08/28/climate/groundwater-drying-climate-change.html>
- Rosen, Z. (2024, September 25). Presidential debate 2024: No Trump bump as Harris captured voters' attention. *CloudResearch*. <https://www.cloudresearch.com/resources/blog/page/2/>
- Roßmann, J., Gummer, T., & Silber, H. (2018). Mitigating satisficing in cognitively demanding grid questions: Evidence from two web-based experiments. *Journal of Survey Statistics and Methodology*, 6(3), 376-400. <https://doi.org/10.1093/jssam/smx020>

- Ryan, T. J. (2018). Data contamination on MTurk. Retrieved from:
<https://timryan.web.unc.edu/2018/08/12/data-contamination-on-mturk/>
- Salerno, J. M., Kulak, K., Smalarz, L., Eerdmans, R. E., Lawrence, M. L., & Dao, T. (2023). The role of social desirability and establishing nonracist credentials on mock juror decisions about Black defendants. *Law and Human Behavior*, 47(1), 100.
- Salk, R. H., Hyde, J. S., & Abramson, L. Y. (2017). Gender differences in depression in representative national samples: Meta-analyses of diagnoses and symptoms. *Psychological Bulletin*, 143(8), 783-822. <https://doi.org/10.1037/bul0000102>
- Savin-Williams, R. C., & Joyner, K. (2014). The dubious assessment of gay, lesbian, and bisexual adolescents of add health. *Archives of Sexual Behavior*, 43(3), 413–422. <https://doi.org/10.1007/s10508-013-0219-5>
- Schachter, S. (1959). *The psychology of affiliation: Experimental studies of the sources of gregariousness*. Stanford University Press
- Schroeder, J., Risen, J. L., Gino, F., & Norton, M. I. (2019). Handshaking promotes deal-making by signaling cooperative intent. *Journal of Personality and Social Psychology*, 116(5), 743. <https://doi.org/10.1037/psp0000117>
- 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(3), 515. <https://doi.org/10.1037/0022-3514.51.3.515>
- Shamon, H., & Berning, C. C. (2020). Attention check items and instructions in online surveys with incentivized and non-incentivized samples: Boon or bane for data quality?. *Survey Research Methods*, 14(1), 55-77. <https://doi.org/10.18148/srm/2020.v14i1.7374>
- Sharman, R. L. (2006). *The tenants of East Harlem*. University of California Press.
- Shelton, J. N., West, T. V., & Trail, T. E. (2010). Concerns about appearing prejudiced: Implications for anxiety during daily interracial interactions. *Group Processes & Intergroup Relations*, 13(3), 329-344. <https://doi.org/10.1177/1368430209344869>

- Siena College Research Institute. (2024, June 27). Double-haters may decide the Presidency. https://scri.siena.edu/wp-content/uploads/2024/06/doubleandlovehaters_6_25.pdf
- Silber, H., Roßmann, J., & Gummer, T. (2022). The issue of noncompliance in attention check questions: False positives in instructed response items. *Field Methods*, 34(4), 346–360. <https://doi.org/10.1177/1525822X221115830>
- Silvia, P. J. (2014). *Write it up: Practical strategies for writing and publishing journal articles*. American Psychological Association.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, 22(11), 1359–1366. <https://doi.org/10.1177/0956797611417632>
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21-word solution. Available at SSRN 2160588. <https://doi.org/10.1177/0956797611417632>
- Simms, L. J., Zelazny, K., Williams, T. F., & Berenstein, L. (2019). Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychological Assessment*, 31(6), 550–560. <https://doi.org/10.1037/pas0000648>
- Simon, S., Moss, A. J., & O'Brien, L. T. (2016). Pick your perspective: Racial group membership and judgments of intent, harm, and discrimination. *Group Processes & Intergroup Relations*, 22(2), 215–232. <https://doi.org/10.1177/1368430217735576>
- Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3, 271–295.
- Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, 166(10), 1092–1097. doi:10.1001/archinte.166.10.1092
- Stagnaro, M. N., Druckman, J., Berinsky, A. J., Arechar, A. A., Willer, R., & Rand, D. (2024). Representativeness versus attentiveness: a comparison across nine online survey samples. *PsyArXiv*, 22.

- Stern, J., Krämer, M. D., Schumacher, A., MacDonald, G., & Richter, D. (2024). Differences between lifelong singles and ever-partnered individuals in Big Five personality traits and life satisfaction. *Psychological Science*, 35(12), 1364-1381.
<https://doi.org/10.1177/09567976241286865>
- Stewart, N., Chandler, J., & Paolacci, G. (2017). Crowdsourcing samples in cognitive science. *Trends in cognitive sciences*, 21(10), 736-748.
- Strickhouser, J. E., Zell, E., & Krizan, Z. (2017). Does personality predict health and well-being? A metasynthesis. *Health psychology*, 36(8), 797.
- Sun, J., Neufeld, B., Snelgrove, P., & Vazire, S. (2022). Personality evaluated: What do people most like and dislike about themselves and their friends?. *Journal of Personality and Social Psychology*, 122(4), 731.
- Sword, H. (2012, October 31). Beware of nominalizations (aka Zombie nouns) - Helen Sword. YouTube. <https://www.youtube.com/watch?v=dNtUJHgcPQ>
- TESS (2024). Time-sharing Experiments for the Social Sciences.
<https://www.tessexperiments.org/>
- Thielmann, I., Akrami, N., Babarović, I., Belsch, A., Bergh, R., Chirumbolo, A., Čolović, P., de Vries, R. E., Fostán, D., Latorova, M., Gnisci, A., Heydasch, T., Hilbig, B. E., Hsu, K.-Y., Izdebski, P., Leone, L., Marcus, B., Michels, M., Perugini, M., ... Lee, K. (2022). The HEXACO-100 across 16 languages: A large-scale test of measurement invariance. *Journal of Personality Assessment*, 104(6), 760-776.
<https://doi.org/10.1080/00223891.2019.1614011>
- Thielmann, I., Moshagen, M., Hilbig, B., & Zettler, I. (2022). On the comparability of basic personality models: Meta-analytic correspondence, scope, and orthogonality of the Big Five and HEXACO dimensions. *European Journal of Personality*, 36(6), 870-900.
- Thorndike F. P., Carlbring P., Smyth F. L., Magee J. C., Gonder-Frederick L., Ost L.-G., Ritterband L. M. (2009). Web-based measurement: Effect of completing single or

- multiple items per webpage. *Computers in Human Behavior*, 25, 393–401.
<https://doi.org/10.1016/j.chb.2008.05.006>
- Todd, A. R., Bodenhausen, G. V., Richeson, J. A., & Galinsky, A. D. (2012). Perspective taking combats automatic expressions of racial bias. *Journal of Personality and Social Psychology*, 100(6), 1027-1042.
- Todd, A. R., Bodenhausen, G. V., & Galinsky, A. D. (2012). Perspective taking undermines stereotype maintenance processes: Evidence from social memory, behavior explanation, and information solicitation. *Social Cognition*, 30(1), 94–108.
<https://doi.org/10.1521/soco.2012.30.1.94>
- Toepoel V., Das M., van Soest A. (2009). Design of web questionnaires: The effects of the number of items per screen. *Field Methods*, 21, 200–217.
<https://doi.org/10.1093/poq/nfh035>
- Tourangeau, R., Conrad, F. G., & Couper, M. P. (2012). *Measurement of web surveys*. Oxford University Press.
- Tourangeau, R., Couper, M. P., & Conrad, F. G. (2004). Spacing, position, and order: Interpretive heuristics for visual features of survey questions. *Public opinion quarterly*, 68(3), 368-393.
- Tupes, E. C., & Christol, R. J. (1992). Recurrent personality factors based on trait ratings. *Journal of Personality*, 60(2), 225-251.
- U.S. Attorney's Office, District of New Hampshire. (2025, April 15). Eight defendants indicted in international conspiracy to bill \$10 million for fraudulent market survey data. United States Attorney's Office. <https://www.justice.gov/usao-nh/pr/eight-defendants-indicted-international-conspiracy-bill-10-million-fraudulent-market>
- Valkenburg, P. M., Meier, A., & Beyens, I. (2022). Social media use and its impact on adolescent mental health: An umbrella review of the evidence. *Current opinion in psychology*, 44, 58-68.

- Vigil-Colet, A., Lorenzo-Seva, U., & Condon, L. (2008). Development and validation of the statistical anxiety scale. *Psicothema*, 20(1), 174-180.
- Villar, A., Callegaro, M., & Yang, Y. (2013). Where am I? A meta-analysis of experiments on the effects of progress indicators for web surveys. *Social Science Computer Review*, 31(6), 744-762. <https://doi.org/10.1177/0894439313497468>
- Weber, M. (2023). Online Sample Fraud: Causes, Costs & Cures. *Case4Quality*.
<https://case4quality.com/about>
- Weems, G.H., & Onwuegbyzie, A.J. (2001). The impact of midpoint responses and reverse coding on survey data. *Measurement and Evaluation in Counseling and Development*, 34(3), 166–176. <https://doi.org/10.1080/07481756.2002.12069033>
- Wessling, K., Huber, J., & Netzer, O. (2017). MTurk character misrepresentation: Assessment and solutions. *Journal of Consumer Research*, 44(2), 211–230.
<https://doi.org/10.1093/jcr/ucx053>
- Wicherts, J., Klein, R. A., Swaans, S. H., Malmgren, F., Stoevenbelt, A. H., Hartgerink, C. H., ... & Ruffer, F. (2022). Privacy protection in the era of open science.
<https://osf.io/preprints/psych/24zu9>
- Willis, G. B. (2004). *Cognitive interviewing: A tool for improving questionnaire design*. sage publications. Sage.
- Wilmot, M. P., & Ones, D. S. (2019). A century of research on conscientiousness at work. *Proceedings of the National Academy of Sciences*, 116(46), 23004-23010.
<https://doi.org/10.1073/pnas.1908430116>
- Wuttke, A., Aßenmacher, M., Klamm, C., Lang, M. M., Würschinger, Q., & Kreuter, F. (2024). AI Conversational Interviewing: Transforming Surveys with LLMs as Adaptive Interviewers. arXiv preprint arXiv:2410.01824.
- Xiao, Z., Zhou, M. X., Liao, Q. V., Mark, G., Chi, C., Chen, W., & Yang, H. (2020). Tell me about yourself: Using an AI-powered chatbot to conduct conversational surveys with

- open-ended questions. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 27(3), 1-37.
- Yan, T., Keusch, F., & He, L. (2018). The impact of question and scale characteristics on scale direction effects. *Survey Practice*, 11(2). <https://doi.org/10.29115/SP-2018-0008>
- Yin, R. K. (2016). *Qualitative research from start to finish*. Guilford publications.
- Zhang, C., & Conrad, F. (2014). Speeding in Web Surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods*, 8(2), 127–135. <https://doi.org/10.18148/srm/2014.v8i2.5453>
- Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, 111(4), 493-504. <https://doi.org/10.1037/pspa0000056>
- Zinsser, W. K. (2016). *On Writing Well: The Classic Guide to Writing Nonfiction*. Quill/A Harper Collins Books.

Subject Index

- anxiety, 62
 behavioral science, 18, 20
 bell curve, 63, 64
 census matched, 89
 ChatGPT,
 communication
 with participants,
 create a study, 87
 dashboard,
 constant, 60
 construct, 7
 construct validity,
 definition
 conceptual, 61
 operational, 61
 demographic,
 evidence, 15
 experimental
 design,
 falsifiable, 14
 Generalized Anxiety Disorder (GAD-7), 98
 Generalized Anxiety Disorder 7 (GAD-7)
 61, 62
 Google Scholar,
 HEXACO, 16
 histogram, 99
 hypothesis, 15
 imitation,
 Insomnia Severity Index (ISI), 98
 institutional review board (IRB),
 instrument, 61
 lexical approach, 12
 Likert,
 manipulation check
 measurement, 7, 59
 scales of,
 mobile device
 metadata,
Moneyball, 20
 moral reasoning, 85
 normal distribution, 63
 operational definition, 7

outlier
 multivariate,
 parsimony, 14
 participant engagement, 8
 participants,
 Patient Health Questionnaire (PHQ-9), 98
 percentile, 6
 personality, 4
 prediction, 64
 predictive validity,
 Primary Care PTSD Screen for DSM-5, 98
privacy,
 quantitative research, 8
 question bank,
 questionnaire, 98
[QuestionPro](#)
 raw score,
 regression
 response option
 direction,
 response options
 vertical vs. horizontal,
 skew
 positive, 100
 survey
 methodology,
 Ten Item Personality Inventory (TIPI), 4
 Ten Item Personality Inventory (TIPI), 4
 Ten Item Personality Inventory, TIPI, 6
 the Big Five, 13
 inventory, 13
 personality traits, 5
 The Heinz Dilemma,
 theory, 12, 13, 14, 16
 third variable,
 trolley dilemma
 footbridge version,
 variability, 8
 variable, 2, 59
 categorical,
 control, 60
 definition, 60

dependent, 60
independent, 60
measured, 60
variation, 63
video,

visualization, 2
voice recording,
writing
structure,

Preview