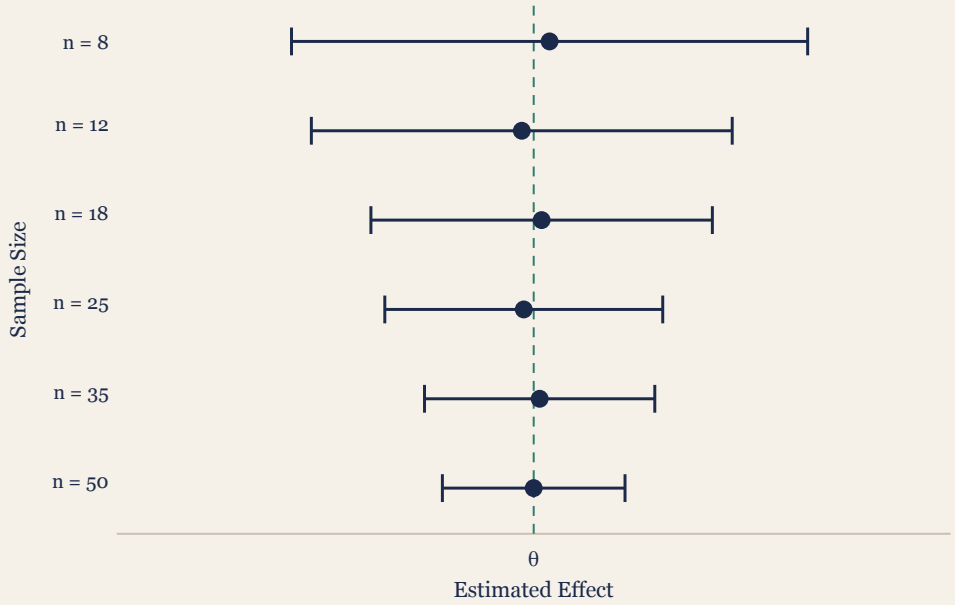
Quantitative Analysis with Small Samples

A Practical Guide for Students and Early-Career Researchers



Confidence Intervals by Sample Size: Uncertainty Narrows as n Grows

Mohammed Ali Sharafuddin
Ahsan Ahmed Jaleel
Meena Madhavan

R · QUARTO · REPRODUCIBLE RESEARCH · APPLIED STATISTICS

Quantitative Analysis with Small Samples

A Practical Guide for Students and Early-Career Researchers

Mohammed Ali Sharafuddin Ahsan Ahmed Jaleel
Meena Madhavan

2026-05-16

Table of contents

Publication Note **12**

Copyright and Licence **14**
Citation and reuse 14
Third-party material 14

Author Contributions **15**
Mohammed Ali Sharafuddin 15
Ahsan Ahmed Jaleel 15
Meena Madhavan 15
Shared responsibility 16
Contributor role summary 16

How to Cite This Book **17**
Suggested citation 17
BibTeX 17
Citation metadata status 17
Versioning 18

Reproducibility and Data Availability **19**
Software environment 19
Data availability 19
Reproducibility expectation 20
Chapter helper functions 20
Archival release 20

Preface **21**
Who This Textbook Is For 21
What You Will Learn 21
Structure of the Textbook 22
Reading Paths 22
Quick Method Selection Guide 23
Conceptual Flow 24
Software and Packages 25
Conventions 25
Acknowledgements 26

Part A: Foundations **27**
In This Part 28

Chapter 1: Why Small-Sample Research Matters **29**
Learning Objectives 29
Why Small Samples Are Often Unavoidable 29
When Large-Sample Approximations Fail 30
Visualising Power Trade-offs 31
Appropriate Methods for Small Samples 33
Example: Comparing Two Small Groups 33
Key Takeaways 35
Self-Assessment Quiz 36
Answers and Explanations 38

Chapter 2: Questions and Outcomes that Fit Small n **41**
Learning Objectives 41
Framing Realistic Research Questions 41
From Objective to Hypothesis 42
Choosing Appropriate Outcomes 42
Effect Sizes and Estimation 44
Example: Outcome Selection in a Pilot Study 45
Research Design Considerations 46
Designing Pilot Studies 47
Key Takeaways 47
Self-Assessment Quiz 47
Answers and Explanations 50

Summary of Part A **53**

Part B: Design, Sampling, and Measurement **54**
In This Part 55

Chapter 3: Sampling Strategies for Small Studies **56**
Learning Objectives 56
The Tension Between Ideal and Feasible Sample Sizes 56
Probability Sampling with Small Samples 56
Sequential and Adaptive Sampling 57
Example: Stratified Sampling Calculation 57
Purposive and Convenience Sampling 59
Quota Sampling 59
Power and Precision with Small Samples 60
Finite Population Correction 60
Example: Power Calculation for a Small Study 61

Sample Size Planning Workflow . . . 62
Justifying Small Sample Sizes . . . 65
Key Takeaways . . . 65
Self-Assessment Quiz . . . 65
Answers and Explanations . . . 68

Chapter 4: Measurement Quality and Scale Development **71**
Learning Objectives . . . 71
The Challenge of Measurement in Small Studies . . . 71
Content and Face Validity . . . 71
Construct and Criterion Validity in Small Samples . . . 73
Reliability, Validity, and Measurement Error . . . 73
Steps for Scale Development with Small Samples . . . 74
Example: Item Analysis for a Pilot Scale . . . 74
Identifying Problematic Items . . . 77
Refining the Scale . . . 78
Qualitative Feedback and Cognitive Interviews . . . 79
Key Takeaways . . . 79
Self-Assessment Quiz . . . 79
Answers and Explanations . . . 82

Chapter 5: Reliability and Measurement Quality for Short Scales **85**
Learning Objectives . . . 85
The Challenge of Short Scales . . . 85
Cronbach's Alpha . . . 85
Interpreting Alpha in Context . . . 86
Example: Cronbach's Alpha for a Short Scale . . . 88
Standard Error of Measurement (SEM) . . . 90
McDonald's Omega . . . 91
Example: McDonald's Omega . . . 92
Split-Half Reliability . . . 93
Example: Split-Half Reliability . . . 93
Worst Split-Half Reliability (Practical Alternative to Revelle's Beta) . . . 95
Polychoric Correlations for Ordinal Items . . . 96
Example: Polychoric Correlations (Conceptual) . . . 97
Reporting Reliability with Small Samples . . . 98
Lab Practical 5.1: Refining a Workplace Resilience Scale . . . 98
Key Takeaways . . . 106
Self-Assessment Quiz . . . 106
Answers and Explanations . . . 109

Chapter 6: Developing Short Scales for Small Samples **112**
Learning Objectives . . . 112
The Scale Development Lifecycle . . . 112

The Iterative Process . . . 112
Special Considerations for n < 50 . . . 120
Minimum Sample Size Requirements . . . 121
What Happens If You Ignore These Requirements? . . . 122
What Should You Do Instead? (For n < 100) . . . 122
Example: Replacing SEM with Composite-Score Analysis . . . 124
Software Will Let You Do Bad Things . . . 125
Recommended Reading (For Future Large-Sample Studies) . . . 125
Bottom Line . . . 126
Example: Documenting a Small-Sample Scale Development . . . 127
Reporting Guidelines for Small-Sample Scale Development . . . 128
Key Takeaways . . . 129
Self-Assessment Quiz . . . 129
Answers and Explanations . . . 132

Chapter 7: Data Screening and Diagnostic Checks **135**
Learning Objectives . . . 135
Why Data Screening Matters More with Small Samples . . . 135
A Practical Screening Workflow . . . 135
Detecting Outliers . . . 136
Example: Outlier Detection with Boxplots and Z-Scores . . . 139
Checking Normality . . . 141
Example: Q-Q Plot for Normality Assessment . . . 142
Linearity and Homoscedasticity in Regression . . . 143
Example: Regression Diagnostics . . . 143
Multicollinearity, Leverage, and Heteroscedasticity . . . 145
Identifying Data Entry Errors . . . 147
Documenting Data Cleaning . . . 147
Key Takeaways . . . 147
Self-Assessment Quiz . . . 148
Answers and Explanations . . . 150

Chapter 8: Handling Missing Data in Small Samples **153**
Learning Objectives . . . 153
The Challenge of Missing Data in Small Samples . . . 153
Types of Missingness . . . 153
Describing Missingness Patterns . . . 154
Example Dataset for Diagnostics . . . 154
Testing the MCAR Assumption . . . 155
Example: Summarising Missing Data . . . 157
Complete-Case (Listwise Deletion) Analysis . . . 159
Mean Imputation (Not Recommended) . . . 162
Last Observation Carried Forward (LOCF) . . . 162
Multiple Imputation (Caution with Small Samples) . . . 163

Example: Multiple Imputation with mice (Caution) . . . 163
Checking Convergence of Multiple Imputation . . . 165
Sensitivity Analyses . . . 165
Preventing Missing Data . . . 165
Key Takeaways . . . 165
Self-Assessment Quiz . . . 165
Answers and Explanations . . . 168

Chapter 9: Assessing Multiple Imputation Quality **170**
Learning Objectives . . . 170
Why Imputation Diagnostics Matter . . . 170
Diagnostic 1: Convergence Checks . . . 170
Diagnostic 2: Imputed vs. Observed Distributions . . . 172
Diagnostic 3: Sensitivity to m (Number of Imputations) . . . 174
Diagnostic 4: Checking Imputation Model Assumptions . . . 177
Diagnostic 5: Fraction of Missing Information (FMI) . . . 179
Example: Full Diagnostic Workflow . . . 180
Red Flags and Troubleshooting . . . 182
Reporting MI Diagnostics . . . 182
Key Takeaways . . . 182
Self-Assessment Quiz . . . 183
Answers and Explanations . . . 186

Summary of Part B **189**

Part C: Analysis Methods **190**
In This Part . . . 191

Chapter 10: Exact Tests and Resampling Methods **192**
Learning Objectives . . . 192
When to Use Exact and Resampling Methods . . . 192
Fisher's Exact Test for 2×2 Tables . . . 193
Example: Fisher's Exact Test . . . 193
When Fisher's Exact Test Is Conservative . . . 195
Unconditional Exact Tests . . . 196
Mid-p Corrections . . . 197
Exact Binomial Test . . . 198
Example: Exact Binomial Test . . . 198
Exact Poisson Test . . . 199
Example: Exact Poisson Test . . . 199
Permutation Tests . . . 200
Example: Permutation Test for Difference in Means . . . 200
Bootstrap Confidence Intervals . . . 202

Example: Bootstrap CI for the Median . . . 202
Key Takeaways . . . 204
Self-Assessment Quiz . . . 204
Answers and Explanations . . . 206

Chapter 11: Nonparametric Rank-Based Methods 209
Learning Objectives . . . 209
When Rank-Based Methods Help . . . 209
Mann–Whitney U Test . . . 209
Effect Sizes Can Be Unstable in Tiny Samples . . . 211
Wilcoxon Signed-Rank Test . . . 212
Kruskal–Wallis and Friedman Tests . . . 213
Rank Correlations . . . 215
Lab Practical 11.1: Sales Performance Analysis . . . 215
Choosing Among Rank-Based Methods . . . 217
Reporting Rank-Based Results . . . 217
Key Takeaways . . . 218
Self-Assessment Quiz . . . 218
Answers and Explanations . . . 220

Chapter 12: Methods for Sparse Counts and Short Time Series 223
Learning Objectives . . . 223
The Challenge of Sparse Counts . . . 223
Choosing a Sparse-Data Method . . . 223
Exact Poisson Test for a Benchmark Rate . . . 224
Comparing Two Sparse Event Rates . . . 225
Bootstrap Forecast Intervals for Short Time Series . . . 226
Zero Inflation and Overdispersion . . . 227
Quasi-Poisson as a Small-Sample Adjustment . . . 228
Negative Binomial and Zero-Inflated Alternatives . . . 228
Reporting Checklist for Sparse-Count Analyses . . . 229
Key Takeaways . . . 229
Self-Assessment Quiz . . . 230
Answers and Explanations . . . 232

Chapter 13: Penalised and Bayesian Regression for Small Samples 234
Learning Objectives . . . 234
The Problem of Sparse Data in Regression . . . 234
Choosing Among Regularisation Strategies . . . 234
Firth-Penalised Logistic Regression . . . 235
Ridge Regression as Shrinkage . . . 236
Choosing the Ridge Penalty with `glmnet` . . . 238
LASSO for Predictor Screening . . . 239
Bayesian Priors as Regularisation . . . 240

Reporting Regularised Models . . . 242
Key Takeaways . . . 242
Self-Assessment Quiz . . . 242
Answers and Explanations . . . 244

Chapter 14: Multi-Criteria Decision Making (MCDM) for Small Sets of Alternatives **245**
Learning Objectives . . . 245
When MCDM Methods Are Appropriate . . . 245
Analytic Hierarchy Process . . . 245
TOPSIS . . . 247
VIKOR and Other MCDM Methods . . . 249
Sensitivity Analysis . . . 250
Key Takeaways . . . 252
Self-Assessment Quiz . . . 252
Answers and Explanations . . . 254

Summary of Part C **255**
Method-Selection Framework . . . 255

Part D: Reporting and Interpretation **258**
In This Part . . . 259

Chapter 15: Effect Sizes and Confidence Intervals over P-Values **260**
Learning Objectives . . . 260
Why P-Values Are Not Enough . . . 260
Common Effect-Size Metrics . . . 260
Mean Differences and Standardised Effects . . . 261
Binary Effects: Odds Ratios, Risk Differences, and NNT . . . 262
Confidence Intervals as the Primary Summary . . . 263
Reporting Effect Sizes . . . 264
Key Takeaways . . . 264
Self-Assessment Quiz . . . 265
Answers and Explanations . . . 266

Chapter 16: Interpreting Non-Significant Results **268**
Learning Objectives . . . 268
What a Non-Significant Result Means . . . 268
Reading the Confidence Interval . . . 268
Power and Minimum Detectable Effects . . . 270
Equivalence and Non-Inferiority . . . 270
Reporting Non-Significant Results . . . 272
Key Takeaways . . . 273

Self-Assessment Quiz . . . 273
Answers and Explanations . . . 275

Chapter 17: Transparent Reporting of Methods and Limitations 278
Learning Objectives . . . 278
The Importance of Transparency . . . 278
Putting the Transparency Pieces Together . . . 278
Documenting Analytic Choices . . . 279
Example: Documenting Analysis Decisions in Code Comments . . . 279
Describing the Sample . . . 281
Example: Sample Characteristics Table . . . 281
Reporting Missing Data . . . 282
Reporting Deviations from Planned Analyses . . . 282
Acknowledging Limitations . . . 282
Handling Multiple Comparisons in Small Samples . . . 283
Pre-Registration for Small-Sample Studies . . . 284
Following Reporting Guidelines . . . 286
Key Takeaways . . . 287
Self-Assessment Quiz . . . 287
Answers and Explanations . . . 289

Chapter 18: Visualising Uncertainty and Presenting Results 291
Learning Objectives . . . 291
The Role of Visualisation in Small-Sample Research . . . 291
Visualising Point Estimates with Confidence Intervals . . . 291
Example: Bar Plot with Error Bars . . . 292
Showing Individual Data Points . . . 293
Example: Dot Plot with Mean and CI . . . 293
Box Plots for Distributional Comparison . . . 295
Example: Box Plot Comparison . . . 295
Visualising Regression Results . . . 296
Example: Scatterplot with Regression Line and CI Band . . . 296
Forest Plots for Several Estimates . . . 298
Raincloud and Half-Eye Plots . . . 299
Presenting Results in Tables . . . 300
Example: Results Summary Table . . . 300
Avoiding Misleading Visualisations . . . 301
Key Takeaways . . . 302
Self-Assessment Quiz . . . 302
Answers and Explanations . . . 304

Summary of Part D 307

Part E: Worked Projects **308**
In This Part . . . 309

Project 1. Evaluating a Marketing Campaign with Ordinal Outcomes **310**
Background . . . 310
Research Question . . . 310
Descriptive Summary . . . 310
Primary Analysis . . . 311
Sensitivity and Subgroup Checks . . . 312
Reporting Summary . . . 313
Extension Task . . . 313

Project 2. Assessing Reliability of a Short Service Quality Scale **314**
Background . . . 314
Item Distributions . . . 314
Reliability Summary . . . 315
Item Diagnostics . . . 316
Branch-Level Check . . . 316
Reporting Summary . . . 316
Extension Task . . . 317

Project 3. Evaluating a Process Improvement Intervention (Paired Design) **318**
Background . . . 318
Descriptive Summary . . . 318
Primary and Sensitivity Analyses . . . 319
Diagnostic Checks . . . 320
Reporting Summary . . . 320
Extension Task . . . 321

Project 4. Evaluating a Reading Intervention in Small Classrooms (Education) **322**
Background . . . 322
Descriptive Summary . . . 322
Primary Analysis . . . 323
Distribution and Classroom Context . . . 324
Reporting Summary . . . 325
Extension Task . . . 325

Project 5. Understanding Intervention Mechanisms: Simple Mediation Analysis **326**
Background . . . 326
Data Summary . . . 326
Regression Paths . . . 327
Bootstrap Indirect Effect . . . 328
Reporting Summary . . . 329
Extension Task . . . 329

Summary of Part E **330**

References **331**

Publication Note

Quantitative Analysis with Small Samples is an open textbook for students and early-career researchers who need to design, analyse, interpret, and report small-sample studies with care. The book gives practical guidance on research questions, measurement quality, diagnostic checks, exact and resampling methods, nonparametric procedures, sparse counts, short time-series, penalised and Bayesian regression, MCDM, uncertainty visualisation, and transparent reporting.

> ⚠ Draft Version 0.1
>
> This public preview is a draft version for coauthor feedback and contribution. Content, examples, metadata, and companion materials may change before formal release.

Authors

- Mohammed Ali Sharafuddin, Qasim Ibrahim School of Business, Villa College, Maldives. ORCID: 0000-0001-5247-2964
- Ahsan Ahmed Jaleel, Qasim Ibrahim School of Business, Villa College, Maldives. ORCID: 0009-0004-5576-8700
- Meena Madhavan, Department of Logistics Management, International Maritime College Oman, National University of Science & Technology, Oman. ORCID: 0000-0003-1244-5392

This Quarto project is organised around a public textbook, with companion teaching materials planned as a separate phase:

- **Textbook**: the open textbook, including the main chapters and worked projects
- **Lab Resources**: planned guided practicals and applied exercises for teaching and workshop use
- **Instructor Manual**: planned syllabi, answer keys, grading guidance, and teaching notes
- **Book Slides**: planned lecture slide sets aligned with the textbook chapters

The current publication target is the public textbook volume. The lab resources, instructor manual, and slides are companion outputs that will be developed and released separately.

Please cite the book using the citation information provided in the **How to Cite** section. Licence and reuse terms are provided in the **Copyright and Licence** section. Reproducibility details are provided in the **Reproducibility and Data Availability** section.

Copyright and Licence

Citation and reuse

Readers, teachers, and researchers may reuse the text, examples, figures, and code for academic and educational purposes with proper citation. Code examples should also cite the repository when reused in teaching material, workshops, software notes, or reproducible research projects.

Third-party material

Every effort has been made to use original examples, generated datasets, open-source tools, and properly cited sources. Any third-party material remains subject to the licence terms of the original copyright holder.

Author Contributions

This page records the agreed author roles for the textbook *Quantitative Analysis with Small Samples: A Practical Guide for Students and Early-Career Researchers*.

Mohammed Ali Sharafuddin

Mohammed Ali Sharafuddin (ORCID: 0000-0001-5247-2964) led the textbook concept, overall structure, chapter planning, pedagogical design, Quarto production workflow, reproducibility strategy, R-based examples, applied statistical framing, cover development, and publisher-readiness preparation. He coordinated the open textbook direction, licensing decisions, repository organisation, and final submission planning.

Ahsan Ahmed Jaleel

Ahsan Ahmed Jaleel (ORCID: 0009-0004-5576-8700) contributed mathematical and statistical expertise, particularly in strengthening the accuracy of quantitative explanations, modelling logic, assumptions, and interpretation. His background in mathematics and his current work on gamified and simulation-based learning support the textbook's emphasis on clear learning design, student engagement, and rigorous yet accessible explanation.

Meena Madhavan

Meena Madhavan (ORCID: 0000-0003-1244-5392) contributed management, logistics, maritime, supply chain, and applied business research expertise. Her experience in business sustainability, maritime logistics, digital transformation, Industry 5.0, green supply chain management, blue economy research, and management education supports the applied examples, case orientation, and early-career researcher positioning of the book.

Shared responsibility

All authors are responsible for reviewing the final manuscript before public release, checking their biographical details and affiliations, confirming the licence statement, and approving the final version submitted to the publisher or deposited with DOI and ISBN metadata.

Contributor role summary

Contributor	Main contribution areas
Mohammed Ali Sharafuddin	Conceptualisation, writing, pedagogy, statistical examples, R and Quarto workflow, reproducibility, editing, publication preparation
Ahsan Ahmed Jaleel	Mathematical and statistical review, modelling accuracy, learning design advice, simulation-based learning perspective
Meena Madhavan	Applied management context, logistics and maritime examples, sustainability and digital transformation perspective, business research review

How to Cite This Book

Please cite the textbook and the source repository when using the material in teaching, research, software, or derivative open educational resources.

Suggested citation

Sharafuddin, M. A., Jaleel, A. A., & Madhavan, M. (2026). *Quantitative Analysis with Small Samples: A Practical Guide for Students and Early-Career Researchers*. GitHub repository and open textbook. https://github.com/MohammedAliSharafuddin/smallsamplelab

BibTeX

```
@book{sharafuddin_jaleel_madhavan_2026_smallsamples,
  title     = {Quantitative Analysis with Small Samples: A Practical Guide
   ↪  for Students and Early-Career Researchers},
  author    = {Sharafuddin, Mohammed Ali and Jaleel, Ahsan Ahmed and
   ↪  Madhavan, Meena},
  year      = {2026},
  publisher = {GitHub},
  url       = {https://github.com/MohammedAliSharafuddin/smallsamplelab}
}
```

Citation metadata status

The repository includes `CITATION.cff`, `.zenodo.json`, and `codemeta.json` to support structured citation. Archival identifiers should be added to those files when the publisher-ready release is deposited. If the book is accepted by an institutional press or open textbook platform, replace the repository publisher field with the final publisher name.

Versioning

Citations should include the version number or release date when a specific release is used. For example, cite Version 1.0.0 when the publisher-ready release is archived.

Reproducibility and Data Availability

This textbook was prepared as a reproducible Quarto project. The source files, R scripts, generated data examples, and rendering configuration are maintained in the project repository.

Repository: https://github.com/MohammedAliSharafuddin/smallsamplelab

Software environment

The project uses R, Quarto, Pandoc, and LaTeX for rendering. The main reproducibility files are:

File	Purpose
`setup.R`	Installs or checks required R packages and prepares datasets where needed
`validation_script.R`	Runs project-level validation checks
`renv.lock`	Records the package versions used for reproducible rendering
`_quarto.yml`	Main Quarto book configuration
`references.bib`	Bibliographic database
`data/`	Teaching and example datasets
`R/`	Supporting R functions and scripts
`.github/workflows/render-book.yml`	GitHub Actions workflow for rendering the book on hosted infrastructure

Data availability

The datasets used in the textbook are teaching datasets or generated example datasets designed for learning and reproducibility. They are not confidential human-subject datasets and should not be interpreted as real institutional or commercial records unless explicitly stated in a chapter.

Reproducibility expectation

A reader should be able to clone the repository, install the required tools, run `renv::restore()`, run the setup script, and render the book. All figures and tables can be regenerated on a fresh machine using `renv::restore()` followed by `quarto render` or the repository render scripts. Final publisher submission should include rendered HTML, PDF, and DOCX outputs, together with render logs and a release archive.

Chapter helper functions

Some chapters use short helper functions to keep table styling and repeated simulation logic consistent across HTML and PDF output. These helpers are not hidden dependencies: they are stored in the repository under `R/` and are sourced by the relevant chapter setup chunks.

The helper files are:

- `R/chapter_table_helpers.R`: provides `smallsamplelab_apa_table()` for consistent APA-style table rendering in HTML and PDF.
- `R/chapter_helpers.R`: provides chapter-specific helpers, including `chapter5_measurement_table()`, `chapter6_simulate_scale()`, `chapter6_problem_item_diagnostics()`, `smalln_mcar_table()`, and `chapter10_format_p()`.
- `R/smalln_recipes.R`: provides reusable small-sample analysis recipes, including `mw_test()` and `boot_median_ci()`.

To inspect or reuse them interactively, run:

```
source("R/chapter_table_helpers.R")
source("R/chapter_helpers.R")
source("R/smalln_recipes.R")
```

Archival release

For formal citation, the publisher-ready version should be archived as Version 1.0.0 and deposited in a DOI-issuing repository. The DOI, ISBN, release date, and final licence should then be inserted into this page, `CITATION.cff`, `.zenodo.json`, `codemeta.json`, and the README.

Preface

This textbook addresses a common challenge in applied research: how to conduct rigorous quantitative analysis when sample sizes are small. Whether you are studying remote communities, rare clinical conditions, pilot educational programmes, or exploratory projects with limited resources, the principles and tools in this guide will help you make sound inferences from modest datasets.

Who This Textbook Is For

This textbook is written for undergraduates, taught masters students, and early-career PhD researchers in **social sciences, health sciences, business, and education** who regularly work with samples of approximately 10 to 100 observations. It is particularly relevant for:

- Researchers conducting studies in Small Island Developing States (SIDS) and similar resource-constrained contexts
- Educational practitioners evaluating classroom interventions with small class sizes
- Health researchers studying rare conditions or conducting pilot clinical trials
- Business analysts testing new strategies in small markets or with limited customer bases
- Social scientists conducting community-based participatory research

What You Will Learn

You will learn to:

- Recognise when small-sample methods are necessary and appropriate.
- Apply exact tests, resampling methods, and rank-based procedures.
- Fit regression models with penalised or Bayesian techniques when classical approaches fail.
- Use multi-criteria decision-making (MCDM) tools for structured evaluation with limited cases.
- Report results transparently, with appropriate uncertainty quantification.

All methods are implemented in R using a curated set of packages. Every code example is designed to run cleanly in a fresh R session, and datasets are small enough to inspect and understand directly.

All figures and tables are intended to be regenerated on a fresh machine by running `renv::restore()` followed by the appropriate Quarto render command from the project root. Helper functions used for repeated formatting or simulation tasks are stored in the `R/` directory so they can be inspected and sourced directly.

Structure of the Textbook

Part A: Foundations introduces the rationale for small-sample research and how to frame research questions that suit limited data.

Part B: Design, Sampling, and Measurement covers sampling strategies, measurement quality, short-scale reliability and development, data screening, and missing-data handling.

Part C: Analysis Methods presents the core toolkit: exact and resampling tests, nonparametric methods, sparse-count and short-time-series methods, penalised and Bayesian regression, and multi-criteria decision-making (MCDM).

Part D: Reporting and Interpretation discusses how to communicate findings, handle uncertainty, and document analytic choices.

Part E: Worked Projects offers complete case studies that integrate multiple methods from earlier chapters.

Planned companion volumes will provide guided lab practicals, instructor-only teaching resources, and chapter-aligned slides. These materials are being developed as a separate release so the textbook can stand on its own as the primary publication.

Self-assessment questions appear in the textbook so students can check their understanding while reading. Full answer keys and suggested grading guidance are planned for the companion Instructor Manual; students should attempt the questions before consulting those materials.

Reading Paths

- **Linear path**: Work through Parts A -> B -> C -> D -> E sequentially for comprehensive coverage.
- **Methods-focused path**: Start with Part B if you already understand the practical constraints of small-sample research.
- **Project-based path**: Begin with Part E and refer back to the earlier chapters when you need method details.

Quick Method Selection Guide

This table is a starting point, not a substitute for design judgement. Use it to locate the most relevant chapter, then check the assumptions, sample-size cautions, and reporting guidance in that chapter before analysing your data.

Research situation	Typical small-n setting	Recommended starting point	Chapter
Planning a study with limited feasible recruitment	n is constrained by budget, access, or population size	Frame one primary question, identify the minimum detectable effect, and separate feasibility aims from confirmatory claims	Chapter 2
Sampling from a known small population	A sampling frame exists but the accessible population is small	Use stratified or finite-population planning when probability sampling is feasible	Chapter 3
Developing or revising a short scale	Early pilot samples of about 5 to 30 participants	Prioritise cognitive interviews, content validity, item diagnostics, and cautious reliability estimates	Chapters 4 to 6
Screening data before analysis	A few unusual cases could affect means, correlations, or regression slopes	Use visual checks, outlier diagnostics, multicollinearity screens, and sensitivity analyses	Chapter 7
Handling missing observations	Missingness is visible but the mechanism is uncertain	Diagnose patterns first; use complete-case analysis only when defensible; use MI cautiously and check diagnostics	Chapters 8 and 9
Comparing small binary or count outcomes	Sparse 2x2 tables, rare events, or benchmark rates	Prefer exact tests and exact intervals before relying on large-sample approximations	Chapters 10 and 12

Research situation	Typical small-n setting	Recommended starting point	Chapter
Comparing skewed or ordinal outcomes	Continuous assumptions are doubtful or the outcome is ordinal	Use rank-based tests with effect sizes and shape-aware interpretation	Chapter 11
Regression with separation or many predictors	Logistic separation, unstable slopes, or p approaching n	Consider Firth logistic regression, ridge/LASSO, or clearly justified Bayesian regularisation	Chapter 13
Choosing among fixed alternatives	Few options must be ranked across multiple criteria	Use AHP, TOPSIS, or VIKOR as transparent decision tools, not as inferential tests	Chapter 14
Reporting a small-sample result	p-values alone do not answer the substantive question	Lead with estimates, confidence intervals, effect sizes, practical thresholds, and uncertainty	Chapters 15 and 16
Writing and presenting the final report	Analytic choices and limitations need to be auditable	Document deviations, multiple comparisons, limitations, and uncertainty visualisations	Chapters 17 and 18

Conceptual Flow

- Part A explains why small-sample research requires different design decisions.
- Part B shows how sampling, measurement, and missing-data choices shape the data before analysis begins.
- Part C introduces the main analytic toolkit.
- Part D focuses on interpretation, reporting, and visual communication.
- Part E integrates the earlier material in full worked projects.

Software and Packages

All analyses use R (version 4.3 or later) and Quarto for reproducible reporting. Core packages include:

- **tidyverse** for data manipulation and visualisation
- **rstatix** for common statistical tests with tidy output
- **boot** for bootstrap resampling
- **exact2x2** for exact tests on 2x2 tables
- **logistf** for Firth-penalised logistic regression
- **glmnet** for ridge, lasso, and elastic net regression
- **mediation** for simple mediation analysis with bootstrap confidence intervals
- **gt** for publication-ready tables
- **performance** for model diagnostics
- **psych** for reliability and factor analysis
- **DescTools** for descriptive and supporting inferential functions
- **MASS** for negative binomial regression
- **scales** for axis labels and percentage formatting
- **naniar** for missing-data summaries and MCAR checks
- **patchwork** for combining plots
- **ggdist** (optional) for raincloud and half-eye uncertainty plots
- **brms** (optional) for Bayesian regression with Stan

```
install.packages(c(
  "tidyverse", "rstatix", "boot", "exact2x2", "logistf", "glmnet",
  "mediation", "gt", "performance", "psych", "DescTools", "MASS",
   ↪  "scales",
  "naniar", "patchwork", "ggdist", "brms"
))
```

Conventions

- British English spelling and punctuation are used throughout.
- Code chunks include `library()` calls so each example can be run independently.
- Random number generation uses `set.seed(2025)` for reproducibility.
- Figures, tables, and worked examples are numbered within chapter.
- File paths are relative to the project root.

Acknowledgements

This textbook draws on the work of many contributors to small-sample methodology, including Van de Schoot and Miočević (2020), Davison and Hinkley (1997), Good (2005), Conover (1999), Firth (1993), Harrell (2015), Hosmer, Lemeshow, and Sturdivant (2013), and Shan (2018).

You are encouraged to work through the chapters in order, running the code examples in your own R environment. The datasets and helper functions referenced in the text are provided in the `data/` and `R/` directories of this project.

Part A: Foundations

Part A establishes the logic for the rest of the book. Chapter 1 explains why small samples are common, where large-sample approximations can mislead, and why small datasets should be analysed on their own terms rather than treated as a deficiency. Chapter 2 turns that foundation into study-design guidance by showing how to frame focused questions, choose informative outcomes, link objectives to hypotheses, and scale claims to what limited data can realistically support.

In This Part

- Chapter 1: Why Small-Sample Research Matters
- Chapter 2: Questions and Outcomes that Fit Small n
- Summary of Part A

Chapter 1: Why Small-Sample Research Matters

Small samples appear routinely across many substantive disciplines: in clinical piloting, ethnographic research, specialised occupational settings, and programme evaluations. In each of these contexts, the productive response to limited n is careful reasoning about power, uncertainty, and method choice rather than apology or retreat. It explains why large-sample approximations can mislead when data are modest, shows how sample size changes what can realistically be detected, and introduces the families of methods that remain useful when information is scarce. The broader aim is to set up the rest of the book: choosing analyses that fit the data you actually have, rather than the data you wish you had.

Learning Objectives

By the end of this chapter, you will be able to explain why small samples are common in applied research, identify where large-sample approximations become unreliable, construct and interpret power curves for different effect sizes, and select methods that fit the data and question at hand rather than the sample size you might have preferred.

Why Small Samples Are Often Unavoidable

Many textbooks assume that researchers can collect hundreds or thousands of observations. In practice, however, numerous research contexts yield small samples. Clinical studies of rare diseases, evaluations of pilot programmes, classroom-based educational interventions, community-based participatory research, and studies in Small Island Developing States (SIDS) often involve fewer than 100 participants. Resource constraints, logistical barriers, and ethical considerations (such as minimising burden on vulnerable populations) make small samples the norm rather than the exception.

For this book, a "small sample" usually means a dataset where routine large-sample approximations cannot be taken for granted: often $n \leq 50$ per group for two-group comparisons, $n \leq 100$ total for simple regression, or fewer than about 20 outcome events for binary or count models. These are working boundaries rather than universal cut-offs. A sample of 80 can still be small for a multivariable logistic model, while a sample of 24 paired observations may be informative for a tightly controlled within-person comparison.

In health sciences, rare-disease trials, feasibility studies, and single-site hospital evaluations may each involve only a few dozen participants. In education, a classroom-based intervention may be tested in one class of 15 to 25 students, while in business and the social sciences the accessible population may be small from the outset, as in niche-market A/B tests, studies of remote communities, or work with specialised occupational groups. Across these settings, the statistical problem runs deeper than a shortfall in recruitment: the population available for study may itself be limited.

Despite their ubiquity, small samples are often treated as deficient or temporary. In practice, this apologetic framing appears when authors describe modest n as a weakness by default, or when reviewers demand larger samples without asking whether a larger pool actually exists, whether recruitment would be ethical, or whether the research question is already well matched to a small but carefully analysed dataset. That response is methodologically unwarranted because sample size is not a free-floating quality marker: its adequacy depends on the population, the effect size, the design, and the inferential goal. In some settings, focused questions about a single outcome or a few key comparisons can often be addressed with modest samples, and even small studies can be informative when effects are large and variability is low. This mindset overlooks the fact that many important questions can only be addressed with small datasets. Rather than apologising, researchers should select methods that are appropriate for the sample size at hand.

When Large-Sample Approximations Fail

Classical parametric tests (t-tests, ANOVA, standard logistic regression) rely on asymptotic theory. They assume that sampling distributions approximate normality as sample size increases. With small samples, these approximations can be inaccurate. P-values may be misleading, confidence intervals may have poor coverage, and maximum likelihood estimates may be unstable or even undefined (for example, in logistic regression with separation).

Small samples also amplify the impact of outliers and violations of distributional assumptions. A single extreme value can dominate a mean or distort a regression slope. Skewed or heavy-tailed distributions, which cause few problems in large samples, become serious concerns when n is small. When n is limited, researchers should also select outcome measures carefully, because continuous or ordinal outcomes usually preserve more information per observation than coarse binary outcomes.

The sample mean remains meaningful, but the reference distribution used by a familiar test may no longer match the actual sampling behaviour well enough. The simulation below shows the point using a one-sample t-test at a nominal two-sided $\alpha = 0.05$ when observations come from a strongly right-skewed distribution with true mean zero. The test is still centred on the correct null value, but the Type I error rate is higher than the nominal 5% level at small n.

Table 0.1: Simulated Type I error for a one-sample t-test under strong right skew.

Sample size	Nominal alpha	Simulated Type I error
5	0.05	0.120
10	0.05	0.103
20	0.05	0.079
30	0.05	0.075

Interpretation: This simulation is deliberately simple, but it gives a concrete reason for caution. At $n = 5$ or 10, a nominal 5% t-test rejects too often under this skewed null distribution. The error rate moves closer to the target as n increases, but the improvement is gradual rather than automatic.

Visualising Power Trade-offs

Even modest reductions in sample size can have a dramatic impact on statistical power. Figure 1.1 uses base R's `power.t.test()` function to illustrate how power declines as per-group sample size falls from 60 to 10, shown separately for small, medium, and large effects (Cohen's $d = 0.3$, 0.5, and 0.8).

To calculate one point on the figure, you supply the per-group sample size `n`, the expected mean difference `delta`, the within-group standard deviation `sd`, the significance level `sig.level`, and the test design through `type = "two.sample"`. For example, `power.t.test(n = 35, delta = 0.5, sd = 1, sig.level = 0.05, type = "two.sample", alternative = "two.sided")$power` returns a value a little above 0.50, meaning that a two-group study with 35 participants per group has only a little better than a fifty-fifty chance of detecting a true effect of $d = 0.5$ at the 5% level. Reading Figure 1.1 in this way helps translate abstract power calculations into concrete design choices.

```
effect_sizes <- c(0.3, 0.5, 0.8)
n_values <- seq(10, 60, by = 5)
alpha <- 0.05

power_grid <- crossing(n = n_values, d = effect_sizes) %>%
  mutate(power = map2_dbl(
    n,
    d,
    ~ power.t.test(
      n = .x,
      delta = .y,
      sd = 1,
      sig.level = alpha,
      type = "two.sample",
```

```
      alternative = "two.sided"
    )$power
  ))

ggplot(power_grid, aes(x = n, y = power, colour = factor(d))) +
  geom_line(linewidth = 1) +
  geom_point() +
  geom_hline(yintercept = 0.80, linetype = "dashed", colour = "grey40") +
  scale_colour_viridis_d(name = "Effect size (d)", option = "D", end =
   ↪  0.85) +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1)) +
  labs(
    x = "Sample size per group",
    y = "Power",
    title = "Statistical power rises with sample size - but slowly at
     ↪  small n",
    subtitle = "Dashed line marks the conventional 80% power threshold"
  ) +
  theme_classic(base_size = 12)
```

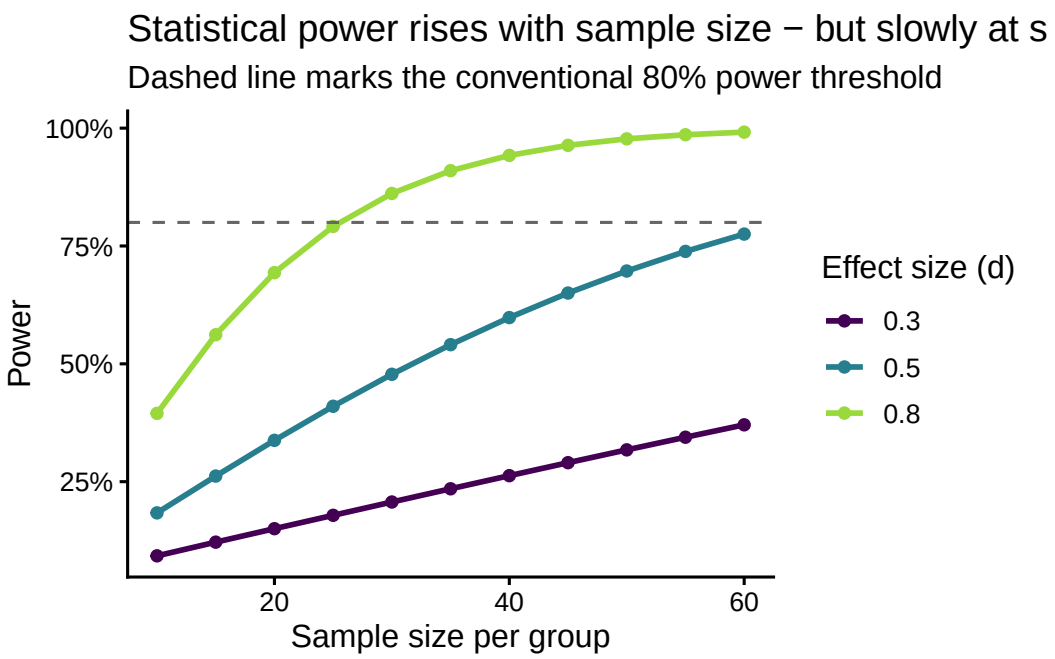


Figure 1.1: Power curves illustrating sensitivity to sample size.

Interpretation

Figure 1.1 shows that with medium effects (about $d = 0.5$), power does not rise above 50% until per-group sample size reaches roughly the mid-30s, and it still remains below the

conventional 80% threshold even at 60 participants per group. Detecting smaller effects (about $d = 0.3$) would require far more observations than are typically feasible in small-sample settings. This visual reinforces the need to report minimum detectable effects and to focus on estimation rather than binary significance testing when n is limited.

Appropriate Methods for Small Samples

When standard asymptotic approximations cannot be trusted, three broad classes of method remain available. Exact tests (such as Fisher's exact test, exact binomial tests, and exact Poisson tests) compute p-values directly from the combinatorial distribution of the data and are especially useful for small discrete datasets. Resampling methods (bootstrap and permutation tests) use the observed data to approximate the sampling distribution, often yielding more accurate inferences than large-sample formulas when an exact calculation is unavailable or when the statistic of interest is more complex than a standard test handles well.

Nonparametric rank-based tests (Mann–Whitney U, Wilcoxon signed-rank, Kruskal–Wallis) make fewer distributional assumptions and are less sensitive to outliers, making them natural choices for ordinal or skewed outcomes. Penalised regression (Firth logistic regression, ridge, LASSO) can stabilise coefficient estimates when events are sparse (that is, when outcome events are rare relative to the number of predictors, as in binary logistic regression with few observed cases of the outcome). Bayesian methods incorporate prior information and quantify uncertainty through posterior distributions, which remain well-defined even when data are limited. In practice, these families complement rather than replace one another: exact tests are often most natural for small discrete tables, resampling methods are flexible for estimation and custom statistics, and penalised or Bayesian models become practical when the research question requires regression rather than a single comparison.

Example: Comparing Two Small Groups

Suppose we wish to compare customer satisfaction scores (on a 1–10 scale) between two service branches, each with only 12 observations. The scores are ordinal and may not be normally distributed.

```
branch_a <- c(7, 8, 6, 7, 9, 8, 7, 6, 8, 7, 9, 8)
branch_b <- c(5, 6, 7, 5, 6, 5, 7, 6, 5, 6, 7, 6)

n1 <- length(branch_a)
n2 <- length(branch_b)
all_scores <- c(branch_a, branch_b)

mw_result <- wilcox.test(branch_a, branch_b, conf.int = TRUE, exact =
 ↪  FALSE)
```

```r
# exact = FALSE avoids tie-related warnings because the satisfaction
 ↪  scores contain repeated values.

tie_counts <- table(all_scores)
tie_term <- sum(tie_counts^3 - tie_counts)
var_u <- n1 * n2 / 12 * ((n1 + n2 + 1) - tie_term / ((n1 + n2) * (n1 + n2 -
 ↪  1)))
z_value <- (as.numeric(mw_result$statistic) - (n1 * n2 / 2)) / sqrt(var_u)
wilcox_r <- abs(z_value) / sqrt(n1 + n2)

summary_table <- tibble(
  branch = c("A", "B"),
  N = c(n1, n2),
  Median = c(
    formatC(median(branch_a), format = "f", digits = 1),
    formatC(median(branch_b), format = "f", digits = 1)
  ),
  IQR = c(
    paste0(
      formatC(as.numeric(quantile(branch_a, 0.25)), format = "f", digits =
       ↪  1),
      "–",
      formatC(as.numeric(quantile(branch_a, 0.75)), format = "f", digits =
       ↪  1)
    ),
    paste0(
      formatC(as.numeric(quantile(branch_b, 0.25)), format = "f", digits =
       ↪  1),
      "–",
      formatC(as.numeric(quantile(branch_b, 0.75)), format = "f", digits =
       ↪  1)
    )
  ),
  Range = c(
    paste0(min(branch_a), "–", max(branch_a)),
    paste0(min(branch_b), "–", max(branch_b))
  )
)

p_text <- if (mw_result$p.value < 0.001) "p < 0.001" else sprintf("p =
 ↪  %.3f", mw_result$p.value)

knitr::kable(
  summary_table,
  align = c("l", "r", "r", "l", "l"),
  booktabs = TRUE
)
```

Table 0.2: Customer satisfaction scores by branch

branch	N	Median	IQR	Range
A	12	7.5	7.0–8.0	6–9
B	12	6.0	5.0–6.2	5–7

```
cat(
  sprintf(
    "\n\n*Note.* Wilcoxon rank-sum test: W = %.0f, %s, Hodges-Lehmann
     ↪  shift = %.2f, standardised r = %.2f. 95%% confidence interval for
     ↪  the location shift: %.2f to %.2f.\n",
    as.numeric(mw_result$statistic),
    p_text,
    unname(mw_result$estimate),
    wilcox_r,
    mw_result$conf.int[1],
    mw_result$conf.int[2]
  )
)
```

Note. Wilcoxon rank-sum test: W = 127, p = 0.001, Hodges–Lehmann shift = 2.00, standardised r = 0.67. 95% confidence interval for the location shift: 1.00 to 2.00.

The Mann–Whitney U test compares the distributions of the two groups without assuming normality. The p-value is the probability of observing a rank difference at least as large as the one obtained, assuming the two groups share the same underlying distribution. Because the test is based on ranks, it is robust to skewness and outliers.

Interpretation

In this example, Branch A scores higher than Branch B, with a Hodges–Lehmann shift estimate of 2 points and a large standardised rank-based effect size (Rosenthal's $r = 0.67$). The small p-value ($p = 0.001$) suggests that the observed rank difference would be unlikely if the two branches had the same underlying distribution. With only 12 observations per branch, the example shows how a rank-based method can still produce a clear, interpretable result when the outcome is ordinal and the effect is sizeable.
That gap between what standard rules of thumb demand and what is actually achievable in constrained settings is precisely the problem this book addresses.

Key Takeaways

Small samples are a common feature of many substantive fields rather than a deficiency to be apologised for, and large-sample approximations can fail when n is modest, leading to inaccurate p-values and confidence intervals. Exact tests, resampling methods, and rank-based procedures

offer valid alternatives that do not require large samples to behave well. In practice, the choice of method should match the research question, outcome type, and available sample size. Small studies are strongest when the analysis is calibrated to the information actually observed and when uncertainty is reported transparently rather than hidden behind a binary significant/non-significant decision. That principle will guide the chapters that follow.

Self-Assessment Quiz

Test your understanding of the key concepts from Chapter 1.

Question 1

A study with n = 12 per group has 25% power to detect d = 0.5. What does this mean?

a) There is a 25% chance the treatment is effective
b) If the true effect is d = 0.5, there is a 25% probability of detecting it ($p < 0.05$)
c) The Type I error rate is 25%
d) 25% of participants will show the effect

Question 2

Why might large-sample approximations fail with n = 15?

a) Computers cannot process small datasets
b) Sampling distributions may not be approximately normal
c) Effect sizes cannot be calculated
d) P-values are always incorrect

Question 3

Which research question is MOST appropriate for n = 20?

a) "What are all factors that predict customer loyalty?" (testing 15 predictors)
b) "Is there a difference in satisfaction between two service approaches?"
c) "How do age, gender, income, education, and occupation interact to predict outcomes?"
d) "Can we build a machine learning model to predict customer behavior?"

Question 4

A pilot study with n = 8 finds a mean difference of 5 points (95% CI: [-2, 12], p = 0.14). The correct interpretation is:

a) There is no effect
b) The effect is exactly 5 points
c) The study is underpowered; effects from -2 to 12 points are plausible
d) The null hypothesis is proven true

Question 5

Which outcome measure provides the MOST statistical information per observation?

a) Binary (pass/fail)
b) Ordinal (grade A-F)
c) Continuous (test score 0-100)
d) All provide equal information

Question 6

With n = 10 per group, which statement about power is TRUE?

a) Power is always 50% regardless of effect size
b) Power increases as the true effect size increases
c) Power is unrelated to sample size
d) Power cannot be calculated for small samples

Question 7

A study finds p = 0.048 with n = 8 per group. Which concern is MOST valid?

a) The result is definitely a false positive
b) With small n, results near the significance threshold should be interpreted cautiously
c) Small samples always produce spurious results
d) The p-value is meaningless with n < 30

Question 8

When is a small sample (n < 30) potentially sufficient?

a) Never—all research requires n >= 100
b) When the effect is very large and variance is low
c) Only for qualitative research
d) When using machine learning methods

Question 9

Which is a legitimate reason for small sample size?

a) The researcher is lazy
b) The population is rare (e.g., a genetic disorder affecting 1 in 100,000)
c) The researcher wants to save time
d) Small samples are always preferable

Question 10

A researcher states: "My study has n = 15, so I'll just use nonparametric tests." What is the problem with this reasoning?

a) Nonparametric tests require n >= 30
b) The choice of test should depend on the data characteristics and research question, not just sample size
c) Nonparametric tests are always inferior
d) Parametric tests always work regardless of assumptions

Answers and Explanations

Question 1

Answer: b)

Explanation: Statistical power is the probability of correctly rejecting a false null hypothesis when a specific effect size exists. With 25% power, there is a 75% chance of a Type II error (failing to detect a real effect of d = 0.5). This concept is directly illustrated in the power curve figure in this chapter, which shows how power declines as sample size decreases.

Question 2

Answer: b)

Explanation: The Central Limit Theorem requires sufficient sample size for sampling distributions to approximate normality. With n = 15, especially if data are skewed or have outliers, parametric test assumptions may be violated. This is why the chapter emphasizes that "large-sample approximations can fail when n is small, leading to inaccurate p-values and confidence intervals."

Question 3

Answer: b)

Explanation: Focused comparisons are feasible with small samples, whereas broad multivariate prediction questions are not. Complex questions like A, C, and D require many more observations to estimate parameters stably and avoid overfitting.

Question 4

Answer: c)

Explanation: The wide confidence interval reflects substantial uncertainty. The study cannot rule out small negative effects (-2) or large positive effects (12). Non-significance with small n indicates insufficient evidence, not absence of effect. This aligns with the chapter's emphasis on focusing "on estimation rather than binary significance testing when n is limited."

Question 5

Answer: c)

Explanation: Continuous measures preserve all variation in the data. Dichotomising or coarsening into categories discards information, reduces statistical power, and limits the ability to detect effects. This is why the chapter notes that, when n is limited, researchers should select outcome measures carefully.

Question 6

Answer: b)

Explanation: Statistical power increases with larger effect sizes, larger sample sizes, and lower variance. Even with n = 10, a very large effect (d = 1.5) might have adequate power, while a small effect (d = 0.2) would not. This is demonstrated in the power curve plot showing different effect sizes (d = 0.3, 0.5, 0.8).

Question 7

Answer: b)

Explanation: P-values near cutoffs (0.05) are highly variable with small samples. A slight change in data or analysis could flip the result. Emphasis should be on effect size magnitude and confidence intervals, not borderline p-values. The chapter warns that "small samples amplify the impact of outliers and violations of distributional assumptions."

Question 8

Answer: b)

Explanation: Small samples can still be informative when effects are large and variability is low. The chapter argues that some important questions can only be addressed with small datasets, provided the method and interpretation are matched to the limited information available.

Question 9

Answer: b)

Explanation: Rare populations, pilot studies, ethical constraints (minimising burden on vulnerable groups), and resource limitations in SIDS contexts are all legitimate reasons for small samples. The chapter explicitly mentions "clinical studies of rare diseases" as one context where "small samples are the norm rather than the exception."

Question 10

Answer: b)

Explanation: Test selection should consider outcome type, distributional properties, and the research question. Small n is one consideration, but not the sole criterion. As the chapter concludes, "The choice of method should match the research question, outcome type, and available sample size."

Chapter 2: Questions and Outcomes that Fit Small n

Small-sample studies are most effective when the research question, outcome, and design are chosen to match the information the data can realistically support. This chapter shows how to narrow broad ideas into answerable questions, how outcome scales affect what can be learned from modest samples, and why estimation often matters more than a binary significant or non-significant result. The goal is practical: to help you design studies that extract the most from limited data and communicate findings at a level the evidence can honestly support.

Learning Objectives

By the end of this chapter, you will be able to distinguish exploratory from confirmatory aims, formulate focused research questions that fit limited data, choose outcome measures that preserve useful information with small n, and calculate the minimum detectable effects implied by a realistic design.

Framing Realistic Research Questions

Small-sample studies work best when the research question is narrow. Questions that ask the data to estimate many predictors, interactions, mediation paths, or measurement parameters such as factor loadings for a long questionnaire usually require far more observations. Focused questions about a single outcome or a few key comparisons are much more realistic with modest samples.

When planning a small-sample study, prioritise clarity and specificity. A question such as "Does a brief reminder intervention improve adherence compared to standard care?" is focused, has a clear comparison, and can be tested in a small randomised trial. A question like "What are all the factors that influence patient adherence?" spreads the available information across too many unknowns.

Similarly, consider whether the study is exploratory or confirmatory. Exploratory studies generate hypotheses, describe patterns, and refine measurement instruments. They can be useful with modest samples, provided the findings are framed as provisional and replication is expected. Because exploratory work often examines several patterns at once, apparent findings may reflect chance, especially when researchers inspect several outcomes or subgroup patterns without adjustment. Confirmatory studies test prespecified hypotheses and therefore require enough power

to support that stronger claim. With small samples, confirmatory aims should be modest and carefully justified.

From Objective to Hypothesis

A small-sample study benefits from a clear hierarchy. In a confirmatory design, that usually means one primary objective, one primary research question, and one primary hypothesis. In an exploratory or pilot design, the hypothesis is often replaced with an estimation or feasibility objective because the data cannot support several formal confirmatory claims well. The objective states what the study is trying to learn, the research question identifies the population, comparison, outcome, and timeframe, and the hypothesis states the expected difference or association on that specific outcome.

For example, a focused small-sample study might use the following sequence:

- **Objective:** To assess whether a brief reminder intervention improves medication adherence over four weeks compared with standard care.
- **Research question:** Among adults attending a primary-care clinic, do participants receiving the reminder intervention have higher four-week adherence scores than those receiving standard care?
- **Confirmatory hypothesis:** Participants assigned to the reminder intervention will have higher mean adherence scores at four weeks than participants assigned to standard care.
- **Exploratory objective:** To estimate the difference in adherence score between groups and assess recruitment, retention, and intervention uptake in preparation for a larger confirmatory trial.

When writing hypotheses for small-sample studies, keep them narrow and defensible. A good small-sample hypothesis names one primary outcome, one main comparison, and a plausible expected pattern. Directional hypotheses are best reserved for situations where prior theory or evidence is strong enough to justify specifying the direction in advance. Avoid omnibus statements that bundle several outcomes, subgroup effects, mediators, and interactions into a single claim. Avoid phrasing hypotheses around achieving statistical significance. The hypothesis should describe the expected substantive pattern, while the analysis later evaluates its uncertainty.

Choosing Appropriate Outcomes

The type of outcome variable influences which methods are feasible and how much information can be extracted from limited data. Binary outcomes (yes/no, success/failure) are common but carry less information per observation than continuous or ordinal measures. If your sample is small, consider whether a continuous or ordinal outcome might capture more variation and yield more precise inferences.

For example, rather than dichotomising patient improvement into “improved” versus “not improved”, use a continuous measure of symptom severity or an ordinal scale with several levels. This preserves information and increases statistical efficiency. When the outcome is inherently binary, such as survival within 30 days, keep it in that form.

Count outcomes (number of adverse events, number of customer complaints) are also informative but may be sparse when samples are small. Exact Poisson tests and negative binomial models can handle low counts, but very sparse data (many zeros, few events) may require careful interpretation or resampling methods.

Outcome Selection Decision Guide

Figure 2.1 turns outcome selection into a sequence of questions. It starts with the construct itself, then separates continuous, count, ordinal, and binary outcomes in the order that preserves the most defensible information from small samples.

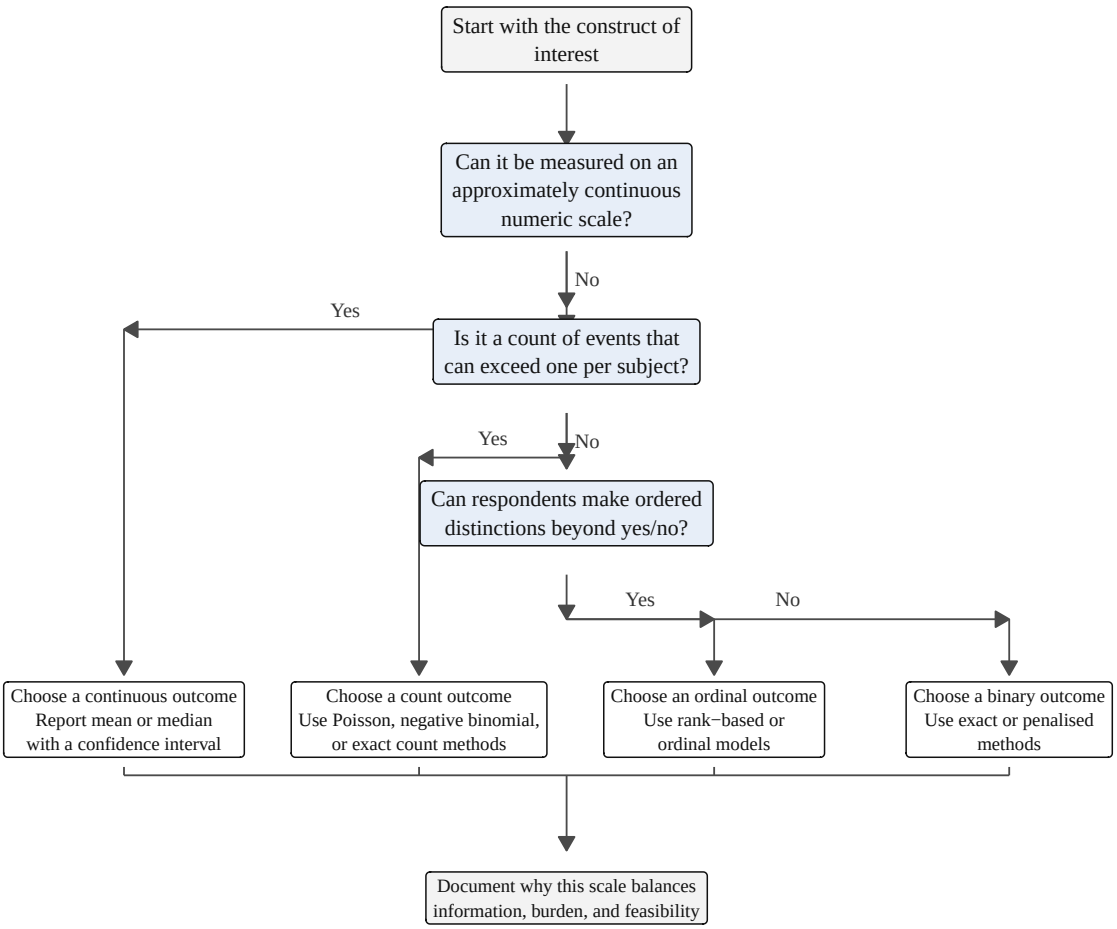


Figure 2.1: Outcome selection guide for small-sample studies.

Read Figure 2.1 from top to bottom. The ambiguous cases usually arise between ordinal and binary coding, or between count outcomes and simpler event/no-event summaries. The guiding principle is to keep the scale that preserves the most defensible information. Binary coding is still appropriate when the construct genuinely has only two meaningful states, or when the substantive decision itself is binary. After choosing the outcome family, explain why that scale balances information, measurement burden, and feasibility for the study.

Effect Sizes and Estimation

In small-sample research, point estimates of effect sizes (differences in means, odds ratios, correlation coefficients) are often more useful than p-values alone. Even when a small sample has limited power, the estimated effect size and its confidence interval indicate the likely magnitude and precision of the effect.

A non-significant result in a small study does not by itself imply that the effect is trivial or absent. It may simply indicate that the data are not precise enough to distinguish a moderate effect from zero with confidence.

When reporting results, emphasise effect sizes and uncertainty intervals. For example, "The mean difference in satisfaction scores was 1.5 points (95% CI: 0.5 to 2.5)" is more informative than "The difference was statistically significant (p = 0.03)". Effect size estimates help readers judge practical importance and facilitate meta-analysis or future sample size planning. When the sample size is fixed in advance, it is also useful to report the minimum detectable effect: the smallest effect your study would be well-positioned to detect under the planned design.

For example, if a two-group study is limited to 15 participants per group and targets 80% power under a two-sided $\alpha = 0.05$, the minimum detectable standardised effect is approximately $d = 1.06$. That means the study is only sensitive to very large differences. Under the same two-sided $\alpha = 0.05$ and 80% power assumptions, detecting a small effect such as $d = 0.2$ would require about 393 participants per group. Thinking in terms of minimum detectable effects helps researchers decide whether a question is realistically answerable with the sample size they can obtain.

As a practical planning check, suppose your budget allows $n = 20$ per group and the planned analysis is a two-sample t-test with two-sided $\alpha = 0.05$ and 80% power. The design is only well positioned to detect about $d = 0.91$ or larger. The next question is substantive rather than computational: would a difference of nearly one pooled standard deviation be the smallest effect worth detecting in your field? If the answer is no, the honest options are to narrow the question, improve measurement precision, use a more efficient paired or stratified design, or frame the study as exploratory rather than confirmatory.

Example: Outcome Selection in a Pilot Study

Suppose you are evaluating a pilot training programme with 18 participants. You have two outcome options: (1) binary pass/fail on a final assessment, or (2) a continuous score (0–100) on the same assessment.

```r
library(dplyr)
library(tibble)

set.seed(2025)
# Inline values quoted in the prose are computed in a separate hidden
 ↪  chunk using the same seed.
n <- 18

scores <- round(rnorm(n, mean = 68, sd = 12))
scores <- pmax(0, pmin(100, scores))

pass_fail <- ifelse(scores >= 60, "Pass", "Fail")

data_pilot <- tibble(
  participant = 1:n,
  score = scores,
  outcome = pass_fail
)

mean_score <- mean(data_pilot$score)
sd_score <- sd(data_pilot$score)
se_score <- sd_score / sqrt(n)
t_crit <- qt(0.975, df = n - 1)
ci_lower <- mean_score - t_crit * se_score
ci_upper <- mean_score + t_crit * se_score
pass_n <- sum(data_pilot$outcome == "Pass")
pass_rate <- pass_n / n
pass_ci <- binom.test(pass_n, n)$conf.int

pilot_summary_table <- tibble(
  Representation = c("Continuous score", "Pass/fail"),
  Estimate = c(
    sprintf("Mean = %.1f; SD = %.1f", mean_score, sd_score),
    sprintf("%d/%d passed (%.1f%%)", pass_n, n, 100 * pass_rate)
  ),
  `Uncertainty / detail` = c(
    sprintf("95%% CI for mean: %.1f to %.1f", ci_lower, ci_upper),
    sprintf("Exact 95%% CI for pass rate: %.1f%% to %.1f%%", 100 *
     ↪  pass_ci[1], 100 * pass_ci[2])
  )
)
```

```r
knitr::kable(
  pilot_summary_table,
  align = c("l", "l", "l"),
  booktabs = TRUE
)
```

Table 0.1: Information retained under two outcome representations.

Representation	Estimate	Uncertainty / detail
Continuous score	Mean = 69.9; SD = 8.4	95% CI for mean: 65.7 to 74.0
Pass/fail	17/18 passed (94.4%)	Exact 95% CI for pass rate: 72.7% to 99.9%

With the continuous score, the sample mean is 69.9 points and the standard deviation is 8.4, with a 95% confidence interval from 65.7 to 74.0. If we dichotomise the same data, the pass rate is 17 out of 18, or 94.4%, with an exact 95% confidence interval from 72.7% to 99.9%. The binary summary still gives a pass rate and its uncertainty, but it no longer shows how far above or below the threshold participants scored.

Interpretation

The continuous outcome lets us estimate average performance and quantify uncertainty directly. If the goal is to understand typical performance rather than only whether participants crossed a cut-point, the continuous measure is more informative.

Research Design Considerations

Small-sample studies benefit from tight experimental control. Paired or matched designs (before–after, crossover, matched-pair comparisons) reduce variability by comparing each unit to itself or a closely matched control. This within-unit comparison can yield precise inferences even when the number of units is small.

Stratification and blocking can also improve efficiency by accounting for known sources of variation. For example, if you are comparing two teaching methods in a small class, stratify by prior achievement level to reduce heterogeneity within each comparison.

Finally, consider sequential or adaptive designs if feasible. Rather than committing to a fixed sample size in advance, you might prespecify an interim review to decide whether recruitment is working as planned, whether variance estimates are much larger than expected, or whether the study should stop early because the signal is already clear. Bayesian methods are well-suited to this style of design because posterior distributions update naturally as data accumulate: the posterior after the first wave becomes the evidence base that is updated when the next wave arrives. These designs still require advance decision rules and transparent reporting so that flexibility remains planned rather than ad hoc.

Designing Pilot Studies

Pilot studies serve specific purposes: assessing feasibility (recruitment rates, attrition, protocol adherence), refining measurement instruments, and estimating variability to inform future sample size calculations. With very small n (often 10–30 participants), focus on collecting process metrics and precision estimates rather than hypothesis testing. Report:

- **Primary feasibility outcomes** (e.g., proportion screened who consent, time to complete assessments).
- **Preliminary effect estimates** with wide confidence intervals, making clear that they are exploratory.
- **Adaptations for the main study**, especially where procedures proved onerous or data quality issues emerged; describe what was changed and why so that reviewers can see how the pilot informed the main design.

In planning terms, choose a pilot sample large enough to detect major logistical problems (often 12–20 per arm is sufficient for estimating key feasibility parameters rather than testing efficacy), prespecify success criteria such as an acceptable recruitment rate, and plan in advance how you will decide whether to proceed to a full trial (Teare et al. 2014).

Key Takeaways

Small-sample studies work best when the question is narrow, the design is realistic, and the outcome preserves as much defensible information as possible. Exploratory aims, continuous or ordinal measures, and efficient designs such as paired or stratified comparisons often make limited data more informative than an overly ambitious confirmatory plan would. Throughout the design process, report effect sizes and confidence intervals alongside power or feasibility considerations so readers can judge what the study could genuinely show.

Self-Assessment Quiz

Test your understanding of the key concepts from Chapter 2.

Question 1

Which research question is better suited to small samples?

a) "What is the relationship between 20 personality traits and job performance?"
b) "Does a brief mindfulness intervention reduce test anxiety compared to control?"
c) "Can we predict customer churn using all available behavioral data?"

d) “How do socioeconomic factors interact to predict health outcomes?”

Question 2

An exploratory study (n=20) finds that meditation reduces anxiety (p=0.04, d=0.7). How should this be framed?

a) “Meditation is proven effective”
b) “Preliminary evidence suggests meditation may reduce anxiety; replication needed”
c) “No conclusions can be drawn from n=20”
d) “The effect is definitely due to chance”

Question 3

A researcher dichotomizes a continuous outcome (0–100 scale) into “high” (70 or above) vs “low” (<70). With n=25, what is the consequence?

a) Power increases because binary outcomes are simpler
b) Power decreases because information is discarded
c) No effect on statistical power
d) Analysis becomes impossible

Question 4

A study aims to detect a “small” effect (d=0.2) with 80% power. Approximately how many participants per group are needed?

a) n=20 per group
b) n=50 per group
c) n=200 per group
d) approximately n=393 per group

Question 5

Which statement about pilot studies is CORRECT?

a) Pilot studies should always test hypotheses
b) Pilot studies assess feasibility and refine procedures
c) Pilot studies require the same sample size as main studies
d) Pilot studies never provide useful effect size estimates

Question 6

A researcher plans a study with n=15 per group but calculates they need n=50 per group for 80% power. What should they do?

a) Proceed exactly as planned and treat the study as confirmatory
b) Reframe the study as exploratory and report the minimum detectable effect for n=15
c) Drop the study because useful information cannot be learned from n=15
d) Keep the original design and add more predictors to recover power

Question 7

Which outcome is LEAST appropriate for n=20?

a) Binary outcome (success/failure)
b) Ordinal outcome (1-7 Likert scale)
c) Continuous outcome (0–100 scale)
d) 50-item questionnaire with subscale factor analysis

Question 8

A study comparing two teaching methods (n=12 per class) finds no significant difference (p=0.18, d=0.45). The conclusion should be:

a) "The two methods are equally effective"
b) "The study found no evidence of a difference, but was underpowered to detect medium effects"
c) "Teaching method has no effect on learning"
d) "The null hypothesis is confirmed"

Question 9

When choosing between a paired and independent-groups design with small samples, which is generally preferable?

a) Always use independent groups—pairing is only for large samples
b) Paired designs reduce within-subject variability and increase power
c) The choice makes no difference statistically
d) Paired designs require larger samples than independent designs

Question 10

A pilot study with n=18 yields a mean difference of 5 points (95% CI: [0.2, 9.8]). What is the appropriate next step?

a) Conclude the intervention is effective and implement widely
b) Use this estimate to plan a fully-powered confirmatory study
c) Abandon the research because the sample was too small
d) Report only the p-value and ignore the confidence interval

Answers and Explanations

Question 1

Answer: b)

Explanation: Focused comparative questions with a single primary outcome are feasible with small samples. Multivariate questions (A, C, D) require large samples to estimate many parameters reliably. The chapter emphasizes: "focused questions about a single outcome or a few key comparisons can often be addressed with modest samples."

Question 2

Answer: b)

Explanation: Exploratory studies with small samples are useful for generating hypotheses, but their findings should be treated as provisional. The chapter emphasizes that such results should be interpreted cautiously, especially because exploratory work can surface patterns that reflect chance and therefore require replication.

Question 3

Answer: b)

Explanation: Dichotomizing continuous variables discards information about the magnitude of differences, reduces statistical power, and can create spurious findings at arbitrary cut-points. The chapter clearly states: "rather than dichotomising patient improvement into 'improved' versus 'not improved', use a continuous measure… This preserves information and increases statistical efficiency."

Question 4

Answer: d)

Explanation: Detecting small effects requires large samples. For d=0.2 with 80% power and alpha = 0.05 (two-tailed), approximately n=393 per group is needed. With n = 15 per group, the chapter shows that a study is only positioned to detect very large effects, approximately d = 1.06 or larger. This aligns with the power curve in Chapter 1, where the d = 0.3 curve remains well below the conventional power target across typical small-sample settings.

Question 5

Answer: b)

Explanation: Pilot studies (typically n=10-30) assess feasibility (recruitment rates, protocol adherence, measurement properties), refine procedures, and provide preliminary effect size estimates for sample size planning, but are not sufficiently large for definitive hypothesis testing. The chapter's "Designing Pilot Studies" section explicitly states: "focus on collecting process metrics and precision estimates rather than hypothesis testing."

Question 6

Answer: b)

Explanation: When the feasible sample size is much smaller than the confirmatory target, the chapter's guidance is to narrow the claim, treat the work as exploratory or pilot-based, and report what effect sizes the study can realistically detect. That combines the framing guidance from the opening section with the emphasis on minimum detectable effects in the estimation section.

Question 7

Answer: d)

Explanation: A 50-item factor analysis asks the data to estimate far too many relationships for n=20. That makes the model unstable and the results difficult to trust. This follows the chapter's broader point that complex multivariate questions are usually unrealistic with very small samples.

Question 8

Answer: b)

Explanation: With small samples, a non-significant result means the study did not provide clear evidence of a difference. An observed effect of d = 0.45 may still be practically meaningful, and the study lacked power to detect it definitively. The chapter emphasizes: "Even when a small sample has limited power, the estimated effect size and its confidence interval indicate the likely magnitude and precision of the effect."

Question 9

Answer: b)

Explanation: Paired designs reduce within-subject variability and increase power. The chapter's "Research Design Considerations" section states: "Paired or matched designs (before–after, crossover, matched-pair comparisons) reduce variability by comparing each unit to itself or a closely matched control. This within-unit comparison can yield precise inferences even when the number of units is small."

Question 10

Answer: b)

Explanation: Use this estimate to plan a fully-powered confirmatory study. Pilot studies provide preliminary effect estimates and variability information needed for sample size planning. The chapter recommends reporting preliminary effect estimates with wide confidence intervals, making clear that they are exploratory, and using pilots to estimate variability for future sample size calculations.

Summary of Part A

Part A established two linked principles. First, small-sample research is common and methodologically legitimate when the design, question, and method are matched to the information available. Second, good small-sample work begins before analysis, with focused objectives, realistic research questions, appropriate outcomes, and claims scaled to what the data can support.

Chapter 1 explained why asymptotic approximations, default large-sample habits, and apologetic framing can all mislead when datasets are modest. Chapter 2 translated that logic into practical design choices by distinguishing exploratory from confirmatory aims, showing how to move from objective to hypothesis, comparing outcome scales, and introducing effect sizes, minimum detectable effects, and pilot-study purposes as planning tools.

The practical lesson is straightforward. Small-sample work is strongest when the question, outcome, design, and analysis are aligned before the data are interpreted. That foundation leads directly to the next part, which turns to the core analytic methods used to answer those questions.

Part B: Design, Sampling, and Measurement

This part addresses the design, measurement, and data-quality decisions that make the later analytic methods in Part C defensible in practice. We cover sampling strategies that maximise information with limited resources, measurement quality and scale development, reliability for short scales, staged short-scale development, data screening and diagnostic checks, and handling missing data transparently.

In This Part

- Chapter 3: Sampling Strategies for Small Studies
- Chapter 4: Measurement Quality and Scale Development
- Chapter 5: Reliability and Measurement Quality for Short Scales
- Chapter 6: Developing Short Scales for Small Samples
- Chapter 7: Data Screening and Diagnostic Checks
- Chapter 8: Handling Missing Data in Small Samples
- Chapter 9: Assessing Multiple Imputation Quality
- Summary of Part B

Chapter 3: Sampling Strategies for Small Studies

Sampling in small studies is less about chasing generic sample-size rules and more about matching design ambitions to the population you can realistically access. This chapter explains how to choose between probability and non-probability sampling, how to justify feasible sample sizes honestly, and how to think about power, precision, and adaptation when recruitment is constrained from the outset.

Learning Objectives

By the end of this chapter, you will be able to explain the trade-offs between probability and purposive sampling, select appropriate sampling methods given realistic resource and population constraints, calculate minimum detectable effects for a planned sample size, and justify sample sizes transparently in research proposals and reports.

The Tension Between Ideal and Feasible Sample Sizes

Many introductory guides rely on heuristics such as n of 30 or more per group for simple group comparisons or 10-15 events per predictor in regression. These are rough planning rules, not universal thresholds. In practice, resource constraints, rare populations, and ethical considerations often make even those targets unattainable. Rather than abandoning research in such contexts, we should adopt methods suited to smaller samples and report findings with appropriate caveats.

Transparent reporting of sampling rationale, achieved sample size, and power or precision estimates helps readers judge the strength of evidence. Researchers should distinguish between studies designed to test specific hypotheses (which require adequate power) and exploratory studies that generate hypotheses or provide preliminary effect estimates (which can proceed with modest samples).

Probability Sampling with Small Samples

Probability sampling (simple random sampling, stratified sampling, cluster sampling) ensures that every unit has a known, non-zero probability of selection. This supports generalisation to

the target population and enables design-based inference. However, probability sampling requires a sampling frame and may be logistically complex or expensive.

With small samples, probability sampling can still be valuable, but estimates will have wide confidence intervals. Stratified sampling, which divides the population into strata and samples proportionally or disproportionately from each, can improve precision by ensuring representation of key subgroups. Probability sampling is most appropriate when a sampling frame is accessible, generalisability to the target population is a genuine aim, and resources permit systematic random selection even if the total sample remains modest.

Sequential and Adaptive Sampling

When recruitment is costly or uncertain, sequential designs allow researchers to review interim results and decide whether to continue sampling. For example, you might pre-specify that recruitment will proceed in waves of five participants, stopping early if credible intervals for the primary outcome are sufficiently narrow or if feasibility metrics (e.g., consent rates) fall below thresholds. Adaptive sampling can also target underrepresented strata after an initial wave, improving balance without committing to a large upfront sample. To keep this defensible, decision rules should be set in advance, error control should be maintained with methods appropriate for small n, and any adaptations should be documented transparently so readers can see how the design evolved.

Sequential or response-adaptive sampling is especially valuable in rare populations, where pausing after each wave prevents over-committing resources if early data already provide actionable evidence.

Example: Stratified Sampling Calculation

Suppose we are surveying employees in a small organisation with 120 total staff: 60 in Department A, 40 in Department B, 20 in Department C. We can afford to survey 30 employees. Proportional stratified sampling ensures each department is represented in proportion to its size.

```
# Population strata
strata <- tibble(
  Department = c("A", "B", "C"),
  Population_N = c(60, 40, 20),
  Proportion = Population_N / sum(Population_N)
)

# Total sample size
total_sample <- 30

# Allocate sample proportionally
strata <- strata %>%
```

```
  mutate(
    Sample_n = round(Proportion * total_sample),
    Sampling_Fraction = Sample_n / Population_N
  )

strata_display_table <- strata %>%
  mutate(
    Proportion = sprintf("%.3f", Proportion),
    Sampling_Fraction = sprintf("%.2f", Sampling_Fraction)
  )

knitr::kable(
  strata_display_table,
  align = c("l", "r", "r", "r", "r"),
  col.names = c("Department", "Population N", "Proportion", "Sample n",
   ↪  "Sampling fraction"),
  booktabs = TRUE
)
```

Table 0.1: Proportional stratified sampling allocation.

Department	Population N	Proportion	Sample n	Sampling fraction
A	60	0.500	15	0.25
B	40	0.333	10	0.25
C	20	0.167	5	0.25

```
cat(
  "\n\n*Note.* The rounded allocation preserves the total sample size of
   ↪  30 and gives each department the same 25% sampling fraction.\n"
)
```

Note. The rounded allocation preserves the total sample size of 30 and gives each department the same 25% sampling fraction.

Interpretation

Proportional allocation ensures that each department contributes to the sample in proportion to its population size. Department A, being the largest, provides 15 respondents, while Department C, the smallest, provides 5. If employees are then selected randomly within each department, this approach yields unbiased estimates for the overall population. If precision for small strata is a concern, disproportionate allocation (oversampling small strata) can be used, though this requires weighting in analysis.

Purposive and Convenience Sampling

Purposive (judgmental) sampling selects units based on researcher judgement of their informativeness or representativeness. Convenience sampling selects units that are easily accessible. Neither method supports probabilistic generalisation, but both are common in small-sample research where probability sampling is infeasible.

Findings from purposive or convenience samples should be interpreted cautiously and presented as preliminary or context-specific. Replication in independent samples strengthens confidence. These approaches are most appropriate when no sampling frame is available, the work is exploratory or pilot-focused, the population is rare or hard to reach, or the available resources are tightly constrained.

Quota Sampling

Quota sampling (a form of purposive sampling) selects units to match known population characteristics (such as age, gender, or occupation distribution). It mimics stratified sampling but without random selection within strata. Quota sampling can improve representativeness compared to convenience sampling, though it remains non-probabilistic. It is most useful when the researcher knows which population characteristics should be balanced but probability sampling is infeasible.

For example, suppose a student satisfaction study can recruit only 30 participants from a programme where the student body is 60% first-year, 25% second-year, and 15% final-year students. A quota plan could recruit 18 first-year, 8 second-year, and 4 final-year students. That improves coverage compared with accepting the first 30 volunteers, but it still does not justify ordinary probability-based confidence intervals unless participants are randomly selected within each quota cell.

Year group	Population share	Quota for n = 30
First year	60%	18
Second year	25%	8
Final year	15%	4

The report should therefore state the quota variables, the target quotas, the achieved quotas, and the recruitment procedure. Findings should be framed as balanced descriptive evidence rather than population estimates with known sampling error.

Snowball sampling is another non-probability approach for hard-to-reach populations. Initial participants refer other eligible participants, creating a sample through social or professional networks. This can be appropriate for rare populations or sensitive topics, but the findings should be interpreted as network-specific because people outside the referral chains had little or no chance of inclusion.

Power and Precision with Small Samples

Statistical power is the probability of detecting a true effect of a given size. With small samples, power is limited, meaning that even if a meaningful effect exists, the study may fail to detect it (high Type II error rate). Researchers should conduct power analyses before data collection to understand what effects are detectable given sample size constraints.

If the achieved sample size is smaller than desired, report the minimum detectable effect (MDE): the smallest effect the study can detect with specified power (typically 80%) and alpha (typically 0.05). This helps readers judge whether the study could have detected effects of practical importance.

Finite Population Correction

When sampling without replacement from a small, known population, the variance of estimates decreases because each sampled unit reduces remaining uncertainty. The finite population correction (FPC) captures that reduction and can lower the required sample size.

In this example, the infinite-population calculation gives a required sample size of 30, but the accessible population is only 120 people. Applying the FPC reduces the required sample size to about 24, as Table 3.2 shows.

```
# Finite population correction example
n_required_infinite <- 30  # From power analysis
N_population <- 120        # Size of accessible population

n_adjusted <- n_required_infinite /
  (1 + (n_required_infinite - 1) / N_population)

fpc_display_table <- tibble(
  `Required n (infinite population)` = n_required_infinite,
  `Accessible population` = N_population,
  `Adjusted n (FPC)` = sprintf("%.2f", n_adjusted),
  Reduction = sprintf("%.2f", n_required_infinite - n_adjusted)
)

knitr::kable(
  fpc_display_table,
  align = c("r", "r", "r", "r"),
  booktabs = TRUE
)
```

Table 0.3: Finite population correction example.

Required n (infinite population)	Accessible population	Adjusted n (FPC)	Reduction
30	120	24.16	5.84

```
cat(
  "\n\n*Note.* The adjusted value is shown to two decimals; in practice,
   ↪  required sample sizes are usually rounded up. Because the FPC
   ↪  reflects a reduction in required precision, rounding to the nearest
   ↪  integer (24) rather than up (25) is the conventional choice here.\n"
)
```

Note. The adjusted value is shown to two decimals; in practice, required sample sizes are usually rounded up. Because the FPC reflects a reduction in required precision, rounding to the nearest integer (24) rather than up (25) is the conventional choice here.

Interpretation

Sampling without replacement from 120 individuals means that a sample of roughly 24 (instead of 30) achieves the same precision. Always report whether you applied the FPC so readers can replicate the calculation.

Example: Power Calculation for a Small Study

We plan a study comparing two groups with n = 12 per group. Under a two-sample t-test, the power to detect a medium effect size (Cohen's d = 0.5) is only about 0.22. Reaching 80% power with this design would require sensitivity to a much larger effect, about d = 1.20, as Table 3.3 shows.

```
# Base R power calculation; setting sd = 1 means delta is on Cohen's d
 ↪  scale.
power_result <- power.t.test(
  n = 12,
  delta = 0.5,
  sd = 1,
  sig.level = 0.05,
  type = "two.sample",
  alternative = "two.sided"
)

# What effect size is detectable with 80% power?
mde_result <- power.t.test(
  n = 12,
  power = 0.80,
```

```
  delta = NULL,
  sd = 1,
  sig.level = 0.05,
  type = "two.sample",
  alternative = "two.sided"
)

power_display_table <- tibble(
  `Per-group n` = 12,
  `Target effect (d)` = sprintf("%.2f", 0.5),
  `Estimated power` = sprintf("%.2f", power_result$power),
  `MDE for 80% power` = sprintf("%.2f", mde_result$delta)
)

knitr::kable(
  power_display_table,
  align = c("r", "r", "r", "r"),
  booktabs = TRUE
)
```

Table 0.4: Power summary for a two-group small study.

Per-group n	Target effect (d)	Estimated power	MDE for 80% power
12	0.50	0.22	1.20

```
cat(
  "\n\n*Note.* Setting `sd = 1` expresses `delta` on Cohen's *d* scale.\n"
)
```

Note. Setting `sd = 1` expresses `delta` on Cohen's *d* scale.

Interpretation

With 12 participants per group, power to detect a medium effect (d = 0.5) is low, at about 22%. To achieve 80% power, the study would need to be sensitive to a much larger effect (about d = 1.2). This illustrates the limitation of small samples for hypothesis testing. If the true effect is small or medium, the study is underpowered. Researchers should acknowledge this limitation and interpret non-significant results cautiously.

Sample Size Planning Workflow

Integrating power analysis into a broader planning conversation prevents unrealistic promises and surfaces design trade-offs early. Use the following workflow whenever you scope a small-sample study:

1. **Clarify the question and estimand.** Identify the exact parameter the study needs to estimate, whether that is a difference in means, an odds ratio, a correlation, or some other target quantity.
2. **Specify tolerable uncertainty.** Define the minimum detectable effect or target confidence-interval width that would make the study actionable in substantive terms.
3. **Map constraints.** Document the recruitment limits, budget, timeline, and ethical restrictions that bound what the design can realistically support.
4. **Select design and analysis.** Choose the planned test or model, decide on one- versus two-sided inference, and note any covariates, matching, or repeated-measures structure.
5. **Compute required n.** Use analytical power formulas, simulation, or resampling as appropriate, and apply finite-population corrections if sampling without replacement from a known population.
6. **Assess feasibility.** Compare the required n with the constraints. If the target is infeasible, revise the design, narrow the claim, or shift the emphasis toward estimation or sequential decision-making.
7. **Document decisions.** Record the assumptions, software, code, and compromises so that readers can see exactly how the final sampling plan was chosen.

Figure 3.1 summarises the planning sequence and the point at which feasibility constraints feed back into design choices.

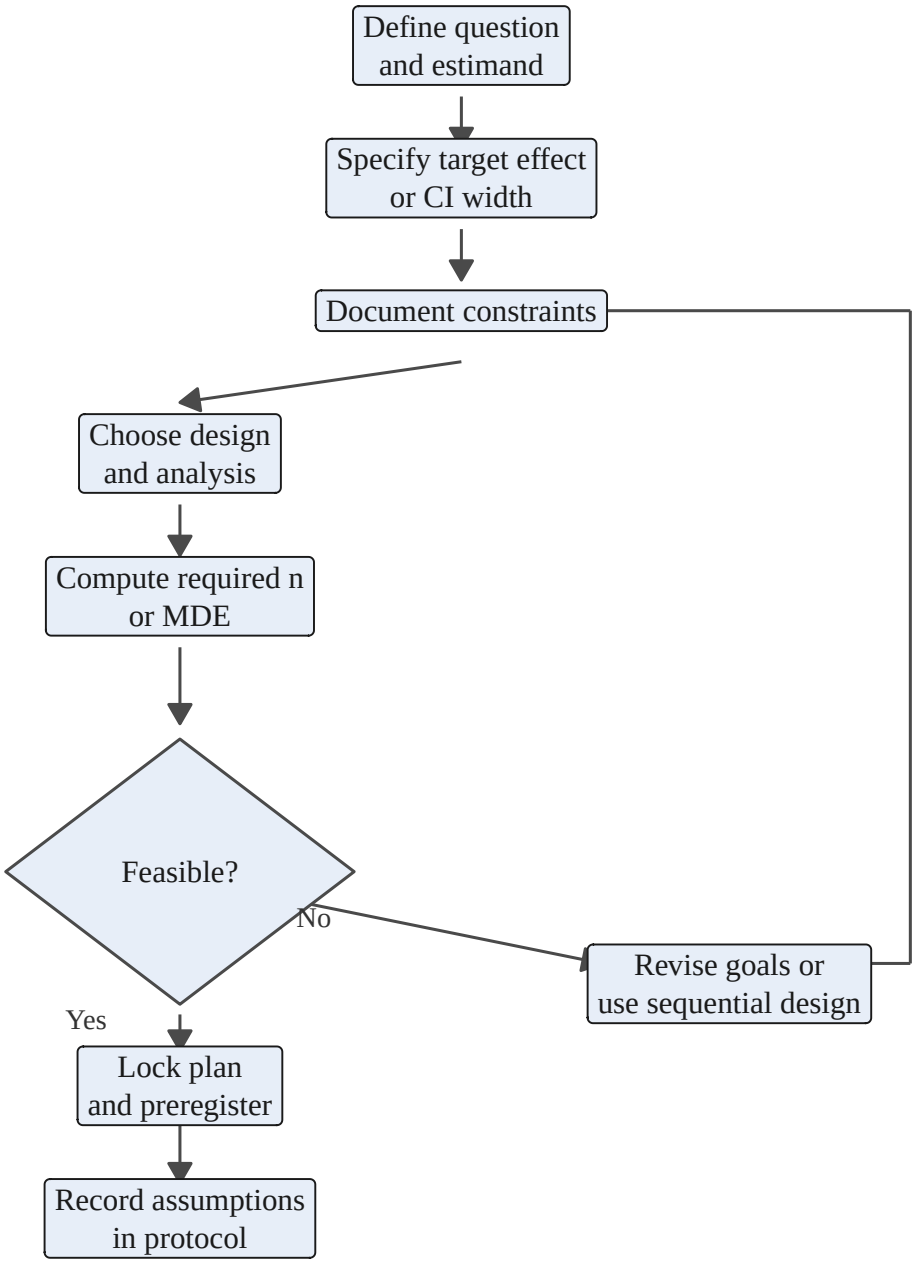


Figure 3.1: Sample size planning workflow for small studies.

Figure 3.1 makes the trade-off loop explicit: if the required sample size exceeds what is feasible, researchers can justify an exploratory framing, add interim analyses, or negotiate for additional resources before data collection begins.

Justifying Small Sample Sizes

When sample sizes are constrained, the researcher's obligation is to be transparent rather than defensive. That means stating clearly both the target and accessible populations, describing the sampling method and the reasoning behind it, and reporting the planned and achieved sample sizes alongside power or precision estimates. If the study cannot detect effects of practical importance, that limitation should be acknowledged directly, and the findings should be framed as preliminary or exploratory when that description is accurate.

Key Takeaways

Sampling in small studies is ultimately about aligning design ambitions with the population you can realistically reach. Probability sampling remains valuable when generalisation is the goal, but purposive, convenience, and quota approaches are often the feasible options in constrained settings and should be described honestly rather than overstated. Across all of these designs, transparent reporting of the sampling rationale, achieved sample, and detectable effect sizes is what allows readers to judge how much weight the findings can bear.

Self-Assessment Quiz

Test your understanding of sampling strategies from Chapter 3.

Question 1

A researcher uses a "rule of thumb" of n=30 per group for all studies. What is the primary problem with this approach?

a) n=30 is always too small
b) Sample size should depend on effect size, power, and research question—not arbitrary rules
c) n=30 is always too large
d) Rules of thumb are always correct

Question 2

Stratified sampling is most useful when:

a) The population is homogeneous
b) You want to ensure representation of key subgroups that differ on the outcome
c) Random selection is impossible
d) Sample size exceeds 1,000

Question 3

Power analysis reveals you need n=50 per group, but only n=20 is feasible. What should you do?

a) Abandon the study
b) Proceed, but report the study as exploratory/pilot and calculate minimum detectable effect (MDE)
c) Proceed and claim the same statistical power
d) Ignore power entirely

Question 4

Which sampling method allows probabilistic generalization to a target population?

a) Convenience sampling
b) Purposive sampling
c) Simple random sampling
d) Snowball sampling

Question 5

Quota sampling differs from stratified sampling in that:

a) It uses random selection within strata
b) It matches population proportions but does not use random selection
c) It requires a sampling frame
d) It is always more accurate

Question 6

A study with n=15 per group has about 25% power to detect d=0.5. The researcher should report:

a) "The study was adequately powered"
b) "The study was underpowered to detect medium effects; only very large effects (around d = 1.06) could be reliably detected"
c) "Power is irrelevant with small samples"
d) "Non-significant results prove no effect exists"

Question 7

The finite population correction (FPC) is relevant when:

a) Sampling with replacement from an infinite population
b) Sampling without replacement from a small, known population (e.g., N=100)
c) Sample size exceeds population size
d) Using convenience sampling

Question 8

Sequential sampling allows researchers to:

a) Collect all data simultaneously
b) Stop early if interim results show sufficient precision or evidence
c) Ignore power analysis
d) Change hypotheses after seeing data

Question 9

A convenience sample from one university is used to test a new teaching method. Which statement is TRUE?

a) Results generalise to all universities
b) Results are context-specific and require replication
c) Convenience sampling is never acceptable
d) Results are as valid as random sampling

Question 10

Minimum Detectable Effect (MDE) refers to:

a) The smallest effect that exists in the population
b) The smallest effect the study can detect with specified power (e.g., 80%)
c) The p-value threshold
d) The confidence interval width

Answers and Explanations

Question 1

Answer: b)

Explanation: Sample size should depend on effect size, power, and research question—not arbitrary rules. A small effect requires larger n; a large effect can be detected with smaller n. The chapter emphasizes: "Rather than abandoning research in such contexts, we should adopt methods suited to smaller samples and report findings with appropriate caveats."

Question 2

Answer: b)

Explanation: Stratified sampling divides the population into strata and ensures each stratum is represented. This improves precision when strata differ on the outcome. The chapter states: "Stratified sampling (dividing the population into strata and sampling proportionally or disproportionally from each) can improve precision by ensuring representation of key subgroups."

Question 3

Answer: b)

Explanation: Proceed, but report the study as exploratory/pilot and calculate minimum detectable effect (MDE). Transparency about power limitations is essential. The chapter recommends: "If the achieved sample size is smaller than desired, report the minimum detectable effect (MDE)."

Question 4

Answer: c)

Explanation: Simple random sampling (and other probability sampling methods) ensures every unit has a known, non-zero probability of selection, supporting generalization. The chapter states: "Probability sampling…ensures that every unit has a known, non-zero probability of selection. This supports generalisation to the target population."

Question 5

Answer: b)

Explanation: Quota sampling matches population proportions but does not use random selection within strata. It "mimics stratified sampling but without random selection within strata." This makes it non-probabilistic.

Question 6

Answer: b)

Explanation: With only about 25% power for d=0.5, the study is underpowered for medium effects. With n=15 per group, only very large effects, around d = 1.06 or larger, are close to the 80% power target. The chapter emphasizes transparent reporting: "Researchers should conduct power analyses before data collection to understand what effects are detectable."

Question 7

Answer: b)

Explanation: FPC adjusts required sample size when sampling without replacement from a small, finite population. The chapter explains: "When sampling without replacement from a small, known population, the variance of estimates decreases…The finite population correction (FPC) adjusts the required sample size accordingly."

Question 8

Answer: b)

Explanation: Sequential sampling allows stopping early based on pre-specified decision rules if interim results show sufficient precision or evidence. The chapter describes: "sequential designs allow researchers to review interim results and decide whether to continue sampling."

Question 9

Answer: b)

Explanation: Convenience samples are context-specific and require replication. The chapter states: "Findings from purposive or convenience samples should be interpreted cautiously and presented as preliminary or context-specific. Replication in independent samples strengthens confidence."

Question 10

Answer: b)

Explanation: MDE is the smallest effect the study can detect with specified power (typically 80%) and alpha (typically 0.05). The chapter defines it: "the minimum detectable effect (MDE): the smallest effect the study can detect with specified power."

Chapter 4: Measurement Quality and Scale Development

Learning Objectives

By the end of this chapter, you will be able to explain the distinctions between content, construct, and criterion validity, pilot and refine short scales with limited samples, compute and interpret basic item-level diagnostics in R, and report measurement evidence transparently when full psychometric validation is not yet feasible.

The Challenge of Measurement in Small Studies

Many small-sample studies rely on brief, custom-developed measurement instruments. Standard scale development protocols (large pilot studies, factor analysis, item response theory) require hundreds of observations. With small samples, researchers must balance the need for reliable, valid measurement with practical constraints.

Short scales (3–5 items) can be internally consistent and valid if items are carefully chosen. Pilot testing with qualitative feedback (cognitive interviews, think-aloud protocols) can identify ambiguous wording, response biases, and cultural appropriateness. Quantitative pilot data (even with n of about 20 to 30) can reveal extreme floor or ceiling effects, items with no variance, and obvious inconsistencies.

Content and Face Validity

Content validity refers to whether items comprehensively and appropriately represent the construct being measured. Face validity refers to whether items appear relevant and appropriate to respondents. Face validity is not a psychometric property in its own right and therefore carries limited evidentiary weight; it should complement, not substitute for, stronger content or construct validity evidence. Both are assessed through expert review and respondent feedback during scale development, before quantitative pilot testing, typically with input from both domain experts and representatives of the target population.

Content Validity Ratio (CVR)

Lawshe's Content Validity Ratio provides a simple index of expert agreement on whether an item is "essential" to a construct. With N experts and n_e rating an item as essential, the CVR is:

$$\text{CVR} = \frac{n_e - N/2}{N/2}$$

CVR ranges from −1 to +1. Positive values indicate majority agreement that the item is essential, but retention thresholds depend on panel size. Under the conventional one-tailed $\alpha = 0.05$ rule, a panel of eight experts requires a CVR of 0.75 or higher, which means at least seven experts must judge the item essential (Lawshe 1975; Ayre and Scally 2014). These critical values correspond to exact binomial probabilities and were re-examined by Ayre and Scally (Ayre and Scally 2014). Use CVR alongside qualitative feedback to decide which items to retain.

```
# CVR example: 8 experts, 6 judge the item essential
n_experts <- 8
n_essential <- 6
cvr <- (n_essential - n_experts / 2) / (n_experts / 2)
cvr
```

```
[1] 0.5
```

Interpretation: A CVR of 0.50 reflects that 6 of 8 experts (75%) rated the item essential. However, for an 8-member panel the usual retention threshold is 0.75, which requires agreement from at least 7 experts (Lawshe 1975; Ayre and Scally 2014). Because 0.50 is below 0.75, this item would not satisfy Lawshe's criterion for statistically significant content validity. Researchers should consult verified critical-value tables when using CVR for item-retention decisions.

Content Validity Index (CVI)

The Content Validity Index is a more descriptive companion to CVR and is common in health, education, and applied measurement work. For an item-level CVI (I-CVI), experts rate each item for relevance, often on a 1–4 scale, and the analyst computes the proportion of experts rating the item as either 3 or 4. A scale-level average CVI (S-CVI/Ave) is then the mean of the item-level CVIs across the candidate item set. CVI does not replace qualitative review, but it gives a transparent summary of expert agreement before a small quantitative pilot begins.

For example, suppose five experts rated six candidate items on a four-point relevance scale. Ratings of 3 or 4 are counted as content-valid endorsements.

Item	Experts rating 3 or 4	I-CVI	Pilot decision
Item 1	5 of 5	1.00	Retain
Item 2	4 of 5	0.80	Retain, check wording
Item 3	3 of 5	0.60	Revise before pilot
Item 4	5 of 5	1.00	Retain
Item 5	2 of 5	0.40	Drop or rewrite
Item 6	4 of 5	0.80	Retain, check redundancy

The S-CVI/Ave for this set is 0.77, calculated as the average of the six I-CVI values. The table is a discussion tool, not a mechanical keep/drop decision rule. Items 3 and 5 need substantive review because several experts did not judge them clearly relevant. Items 2 and 6 may be acceptable, but the qualitative comments should be checked for ambiguity or overlap.

Construct and Criterion Validity in Small Samples

Content validity is only one part of measurement quality. Researchers also need to consider whether scores behave as theory predicts, which is the core of **construct validity**, and whether they relate to a meaningful external benchmark, which is the core of **criterion validity**. In small studies, these checks should usually be modest and pre-specified. Rather than claiming a full psychometric validation from n = 20–40, look for tentative supporting evidence by comparing groups that theory predicts should differ on the construct, examining correlations with one or two closely related measures while reporting effect sizes and confidence intervals, or checking whether scores predict a practically relevant external outcome. Report effect sizes and uncertainty, avoid post-hoc fishing for significant associations, and avoid strong claims from unstable estimates.

Reliability, Validity, and Measurement Error

Reliability concerns **consistency**; validity concerns whether the instrument measures the **intended construct**. A scale can be reliable without being valid. In an early pilot, item statistics and internal consistency estimates are best treated as screening tools that flag obvious weaknesses, not as definitive proof that the scale is ready for high-stakes use. Chapter 5 extends this discussion for short scales.

If items are scored or coded by human raters, agreement among raters becomes part of measurement quality. With two raters, report a chance-corrected agreement statistic such as Cohen's kappa alongside the raw agreement percentage; with more than two raters, use a multi-rater extension such as Fleiss' kappa or an intraclass correlation coefficient when scores are numeric. In small samples these estimates can be unstable, so describe the coding protocol, rater training, and disagreement-resolution process rather than reporting a coefficient alone.

Steps for Scale Development with Small Samples

1. **Define the construct clearly.** Start by specifying what you are measuring and which dimensions or facets belong inside that construct.
2. **Generate candidate items.** Write more items than you expect to keep so that weak or redundant items can be removed later without leaving gaps in content coverage.
3. **Expert review.** Ask domain experts to rate each candidate item for relevance, clarity, and representativeness, and use their feedback to flag wording or content problems early.
4. **Cognitive interviews.** Ask a small number of respondents (often n = 5–10) to complete the scale while thinking aloud so you can hear how they interpret the wording and response options.
5. **Quantitative pilot.** Administer the revised scale to a small sample (often n = 20–40) and compute descriptive item statistics to identify obvious weaknesses.
6. **Item analysis.** Review patterns such as low variance, weak corrected item-total correlations, and floor or ceiling effects to decide which items need revision.
7. **Refine and re-test.** Remove or revise problematic items and, if resources permit, test the revised version again because scale development is iterative rather than strictly linear (DeVellis and Thorpe 2021).

Example: Item Analysis for a Pilot Scale

We pilot a 5-item job satisfaction scale with n = 25 employees. Each item uses a 1–7 Likert response.

```
library(tidyverse)
library(psych)

set.seed(2025)
n <- 25
latent_satisfaction <- rnorm(n)

# Simulated pilot data: 25 respondents, 5 items
# clear_communication, staff_courtesy, and overall_satisfied track the
 ↪  same construct.
# wait_time_ok has a restricted high-end range; problem_resolved
 ↪  discriminates poorly.
lkrt7 <- function(z) {
  as.integer(cut(z, breaks = c(-Inf, -1.5, -0.9, -0.3, 0.3, 0.9, 1.5,
   ↪  Inf), labels = FALSE))
}
pilot_data <- tibble(
  respondent           = 1:n,
  clear_communication  = lkrt7(latent_satisfaction + rnorm(n, 0, 0.8)),
  staff_courtesy       = lkrt7(latent_satisfaction + rnorm(n, 0, 0.8)),
```

```r
  wait_time_ok         = pmax(4L, lkrt7(0.25 * latent_satisfaction + 1.3 +
   ↪  rnorm(n, 0, 0.55))),
  problem_resolved     = lkrt7(0.10 * latent_satisfaction + rnorm(n, 0,
   ↪  1.2)),
  overall_satisfied    = lkrt7(latent_satisfaction + rnorm(n, 0, 0.8))
)

# Item descriptive statistics
item_stats <- pilot_data %>%
  select(clear_communication, staff_courtesy, wait_time_ok,
   ↪  problem_resolved, overall_satisfied) %>%
  summarise(across(everything(), list(
    mean = mean,
    sd = sd,
    min = min,
    max = max
  ))) %>%
  pivot_longer(everything(), names_to = c("item", ".value"), names_pattern
   ↪  = "^(.+)_(mean|sd|min|max)$")

# Inter-item correlations
items_only <- select(pilot_data, clear_communication, staff_courtesy,
 ↪  wait_time_ok, problem_resolved, overall_satisfied)
cor_matrix <- cor(items_only)

# Item-total correlations (corrected for item overlap)
invisible(capture.output(
  alpha_full <- suppressWarnings(psych::alpha(items_only))
))
item_total <- alpha_full$item.stats %>%
  as.data.frame() %>%
  rownames_to_column("item") %>%
  transmute(
    item,
    r_drop = round(r.drop, 2),
    mean = round(mean, 2),
    sd = round(sd, 2)
  )

item_stats_display <- item_stats %>%
  rename(Item = item, Mean = mean, SD = sd, Min = min, Max = max)

cor_display <- round(cor_matrix, 2) %>%
  as.data.frame() %>%
  rownames_to_column("Item")

item_total_display <- item_total %>%
  rename(
    Item = item,
```

```
    `Corrected item-total correlation` = r_drop,
    Mean = mean,
    SD = sd
  )

print(knitr::kable(
  item_stats_display,
  align = c("l", "r", "r", "r", "r"),
  booktabs = TRUE,
  caption = "Item descriptive statistics for the pilot scale."
))
```

Table 0.2: Item descriptive statistics for the pilot scale.

Item	Mean	SD	Min	Max
clear_communication	4.32	1.9732	1	7
staff_courtesy	4.20	1.7078	1	7
wait_time_ok	6.08	0.7594	5	7
problem_resolved	3.60	1.8028	1	7
overall_satisfied	4.12	1.7156	1	7

```
print(knitr::kable(
  cor_display,
  align = c("l", "r", "r", "r", "r", "r"),
  booktabs = TRUE,
  caption = "Inter-item correlation matrix for the pilot scale."
))
```

Table 0.3: Inter-item correlation matrix for the pilot scale.

Item	clear_communication	staff_courtesy	wait_time_ok	problem_resolved	overall_satisfied
clear_communication	1.00	0.34	0.18	-0.20	0.46
staff_courtesy	0.34	1.00	0.12	0.03	0.55
wait_time_ok	0.18	0.12	1.00	0.27	0.41
problem_resolved	-0.20	0.03	0.27	1.00	0.06
overall_satisfied	0.46	0.55	0.41	0.06	1.00

```
print(knitr::kable(
  item_total_display,
  align = c("l", "r", "r", "r"),
  booktabs = TRUE,
  caption = "Corrected item-total correlations for the pilot scale."
))
```

Table 0.4: Corrected item-total correlations for the pilot scale.

Item	Corrected item-total correlation	Mean	SD
clear_communication	0.29	4.32	1.97
staff_courtesy	0.45	4.20	1.71
wait_time_ok	0.38	6.08	0.76
problem_resolved	-0.01	3.60	1.80
overall_satisfied	0.62	4.12	1.72

Reproducibility note: When adapting this code for your own pilot, record your `set.seed()` value and package versions so the same simulated item patterns can be regenerated.

Cronbach's alpha for the initial 5-item scale was 0.540. Because this pilot uses only 25 respondents, the estimate should be treated cautiously: alpha can be unstable with small samples, so in applied work it is sensible either to report confidence intervals or to note the sampling uncertainty around the point estimate.

Interpretation: Look for a broadly coherent pattern rather than demanding that every statistic fall within a rigid threshold. In this pilot, Items 1, 2, and 5 move together reasonably well. Item 3 has a restricted high-end range and a negative corrected item-total correlation, suggesting that it is not functioning like the rest of the scale. Item 4 contributes very little to the total score, which makes it a candidate for revision or removal as well. The corrected item-total correlation (`r.drop`) is especially useful here: values below approximately 0.30 often indicate weak discrimination and deserve closer review, but the threshold is context-dependent. Very short scales may legitimately show lower average inter-item correlations, and substantively important facets should not be dropped solely on statistical grounds.

Identifying Problematic Items

Problematic items usually reveal themselves through a combination of weak descriptive and correlational signals. If an item has very low variance, most respondents are giving essentially the same answer, which often means the wording is too obvious, too extreme, or too narrow. If the corrected item-total correlation is weak, the item may be discriminating poorly or may be tapping a different construct. Floor or ceiling effects limit an item's ability to distinguish among respondents because most answers cluster at one end of the scale. Negative correlations are especially important warning signs because they often indicate incorrect reverse coding or an item working against the rest of the scale. Treat these patterns as heuristics rather than automatic deletion rules: a statistically weak item may still capture an important facet of the construct and be worth revising rather than dropping.

Refining the Scale

Based on item analysis, revise or remove problematic items. For example, if Item 3 shows a ceiling effect and Item 4 has weak item-total correlation, consider removing them. Compute alpha for the revised scale.

```
# Revised scale: keep the three items that cohere best
invisible(capture.output(
  alpha_full <- suppressWarnings(psych::alpha(select(pilot_data,
 ↪  clear_communication, staff_courtesy, wait_time_ok, problem_resolved,
 ↪  overall_satisfied)))
))
revised_items <- select(pilot_data, clear_communication, staff_courtesy,
 ↪  overall_satisfied)
invisible(capture.output(
  alpha_revised <- suppressWarnings(psych::alpha(revised_items))
))

alpha_comparison <- tibble(
  scale = c("Initial 5-item scale", "Revised 3-item scale"),
  alpha = c(alpha_full$total$raw_alpha, alpha_revised$total$raw_alpha),
  average_inter_item_r = c(alpha_full$total$average_r,
   ↪  alpha_revised$total$average_r)
) %>%
  mutate(across(where(is.numeric), ~ round(.x, 3))) %>%
  rename(
    Scale = scale,
    `Cronbach's alpha` = alpha,
    `Average inter-item correlation` = average_inter_item_r
  )

print(knitr::kable(
  alpha_comparison,
  align = c("l", "r", "r"),
  booktabs = TRUE,
  caption = "Internal consistency before and after item refinement."
))
```

Table 0.5: Internal consistency before and after item refinement.

Scale	Cronbach's alpha	Average inter-item correlation
Initial 5-item scale	0.540	0.220
Revised 3-item scale	0.702	0.447

Interpretation: In this example, dropping Items 3 and 4 raises alpha from the initial 5-item version to the revised 3-item version. That does **not** mean researchers should always delete items with low

statistics. Item removal should also reflect the construct definition, expert review, and respondent feedback. If Item 3 captures an important facet of job satisfaction, revising its wording may be preferable to dropping it entirely. For very short scales, it is also useful to inspect the average inter-item correlation rather than relying on alpha alone (DeVellis and Thorpe 2021).

Qualitative Feedback and Cognitive Interviews

With very small samples (n < 20), quantitative item analysis is unreliable. Qualitative methods (cognitive interviews, focus groups) are more informative. Ask respondents:

- What does each item mean to you?
- Were any items confusing, ambiguous, or difficult to answer?
- Are the response options appropriate?
- Are any items culturally inappropriate or offensive?

This feedback can prevent major problems before larger-scale data collection.

Key Takeaways

Measurement quality matters especially in small-sample research because unreliable measures reduce power and can distort substantive conclusions. In practice, strong small-sample scale development depends less on claiming full psychometric validation and more on combining expert review, respondent feedback, and cautious item-level diagnostics to identify obvious weaknesses. Iterative refinement, rather than one definitive pilot, is what gradually improves the scale when resources are limited.

Self-Assessment Quiz

Question 1

What is the primary advantage of using qualitative methods (cognitive interviews) over quantitative methods when pilot testing scales with very small samples (n < 20)?

a) Cognitive interviews provide more statistical power
b) Qualitative feedback can identify ambiguous wording and cultural issues without requiring statistical reliability
c) Quantitative item analysis is too expensive
d) Cognitive interviews automatically calculate Cronbach's alpha

Question 2

In Lawshe's Content Validity Ratio (CVR), if 8 experts are consulted and 6 judge an item as "essential," what is the CVR?

a) 0.25
b) 0.50
c) 0.75
d) 1.00

Question 3

What does a "ceiling effect" in item analysis indicate?

a) Most respondents give the lowest possible response
b) Most respondents give the highest possible response
c) The item has perfect reliability
d) The item correlates negatively with the total score

Question 4

Which corrected item-total correlation threshold typically indicates that an item discriminates poorly and should be considered for removal?

a) Above 0.7
b) Between 0.3 and 0.7
c) Below 0.3
d) Exactly 0.5

Question 5

What does content validity assess?

a) Whether items comprehensively and appropriately represent the construct being measured
b) Whether the scale has high Cronbach's alpha
c) Whether factor analysis confirms a one-dimensional structure
d) Whether the scale predicts future behavior

Question 6

In a small pilot study, what would count as preliminary evidence for construct validity?

a) A single high Cronbach's alpha
b) Scores relate to other variables or known groups in the direction theory predicts
c) Respondents say the scale looks professional
d) A factor analysis with n = 25 produces one factor

Question 7

In pilot testing with small samples (n of about 20 to 30), what is the primary limitation of conducting factor analysis?

a) Factor analysis requires specialised software
b) Factor analysis usually needs much larger samples for stable loadings; n of about 20 to 30 is typically too small
c) Factor analysis only works with 7-point Likert scales
d) Factor analysis cannot handle missing data

Question 8

If an item correlates negatively with the total score and with other items, what is the most likely explanation?

a) The item has high content validity
b) The item may be reverse-coded incorrectly or measuring an opposite construct
c) The sample size is too large
d) The item has a ceiling effect

Question 9

What does criterion validity ask?

a) Whether scale scores align with an external benchmark or relevant outcome
b) Whether experts agree an item is essential
c) Whether all items load on one factor
d) Whether a scale has no missing values

Question 10

In the example of the 5-item job satisfaction scale, Item 3 had a restricted range (responses only from 4–7 on a 1–7 scale). What does this suggest?

a) Item 3 has perfect reliability
b) Item 3 may have a ceiling effect, limiting its ability to differentiate among respondents
c) Item 3 should be kept because high scores are desirable
d) The sample size should be increased to 1,000

Answers and Explanations

Question 1

Answer: b)

Explanation: "With very small samples (n < 20), quantitative item analysis is unreliable. Qualitative methods (cognitive interviews, focus groups) are more informative." Cognitive interviews reveal ambiguous wording and cultural issues without requiring the large samples needed for statistical reliability indices.

Question 2

Answer: b)

Explanation: CVR = (6 - 8/2) / (8/2) = (6 - 4) / 4 = 2/4 = 0.50. That means 75% of experts judged the item essential. However, with N = 8 the usual Lawshe threshold is 0.75, so a CVR of 0.50 would not satisfy the standard retention criterion.

Question 3

Answer: b)

Explanation: "If most responses cluster at the low or high end, the item cannot differentiate among respondents." A ceiling effect occurs when most respondents give the highest possible response, limiting discrimination.

Question 4

Answer: c)

Explanation: "The corrected item-total correlation (r.drop) indicates how well each item correlates with the total score excluding itself. Values below about 0.3 often suggest weak items that deserve closer review." Such items may discriminate poorly, but final decisions should still consider scale length, construct coverage, and respondent feedback.

Question 5

Answer: a)

Explanation: "Content validity refers to whether items comprehensively and appropriately represent the construct being measured." Content validity is assessed through expert review, not statistical tests.

Question 6

Answer: b)

Explanation: Construct validity concerns whether scores behave in ways theory predicts. In small studies, this usually means modest, pre-specified checks such as known-groups comparisons or correlations with closely related measures, reported with effect sizes and uncertainty rather than post-hoc fishing for significant associations.

Question 7

Answer: b)

Explanation: Factor analysis usually needs substantially larger samples than a small pilot can provide. Common rules of thumb suggest about 5 to 10 participants per item or an absolute N of at least 100, although the exact requirement depends on communalities, item quality, and model complexity. A pilot sample of 20 to 30 is therefore usually too small for stable factor loadings, though it can still support preliminary item screening.

Question 8

Answer: b)

Explanation: "If an item correlates negatively with the total or with other items, it may be reverse-coded incorrectly or measuring an opposite construct." Negative correlations suggest coding errors or conceptual misalignment with the scale.

Question 9

Answer: a)

Explanation: Criterion validity asks whether scores relate to an external benchmark, either at the same time (concurrent validity) or later (predictive validity). In small studies, this evidence should be reported cautiously with effect sizes and uncertainty.

Question 10

Answer: b)

Explanation: "Item 3 has a restricted range (4–7), which may indicate a ceiling effect." Restricted ranges (especially at the high end) indicate ceiling effects that limit the item's ability to differentiate among respondents.

Chapter 5: Reliability and Measurement Quality for Short Scales

Learning Objectives

By the end of this chapter, you will be able to explain why reliability estimates behave differently in short scales, distinguish alpha, omega, and split-half coefficients, compute the main internal-consistency diagnostics in R, and report measurement quality with appropriate confidence intervals and small-sample caveats.

The Challenge of Short Scales

Many small-sample studies use brief measurement instruments (3–5 items) to reduce respondent burden. Short scales, however, pose challenges for reliability assessment. Classical reliability indices (Cronbach's alpha, split-half reliability) are attenuated when scales have few items. Moreover, small sample sizes yield imprecise reliability estimates with wide confidence intervals.

Despite these limitations, reliability assessment remains essential. Unreliable measures introduce noise, reducing statistical power and biasing effect estimates. Researchers should report reliability alongside validity evidence and interpret findings cautiously when reliability is low.

Cronbach's Alpha

Cronbach's alpha estimates internal consistency by comparing item variances to total scale variance. It assumes that all items measure a single underlying construct and that item true scores contribute equally to the total score, usually described as tau-equivalence. A stricter parallel-items model also requires equal error variances, whereas a congeneric model allows items to have different loadings. Alpha increases with the number of items and the average inter-item correlation.

Assumptions: Items are at least approximately tau-equivalent. Errors are uncorrelated. Continuous or approximately continuous item responses.

When to use: Multi-item scales (3 or more items), desire for simple internal consistency estimate. Interpret cautiously for short scales and with small samples.

Interpreting Alpha in Context

Traditional thresholds (e.g., $\alpha \geq 0.70$ for research, $\alpha \geq 0.90$ for high-stakes decisions) are **heuristic benchmarks, not universal rules** (DeVellis and Thorpe 2021):

- **Exploratory research / pilot studies**: $\alpha = 0.60$–0.70 acceptable
- **Established scales in research**: $\alpha = 0.70$–0.90 expected
- **High-stakes decisions**: $\alpha > 0.90$ required
- **Very short scales** (3 items): $\alpha = 0.50$–0.65 may be acceptable only when items are conceptually narrow and homogeneous, corrected item-total correlations exceed about 0.30, and the estimate is reported with its confidence interval

The alpha coefficient alone is rarely sufficient for assessing measurement quality. More diagnostic in most small-sample settings are the item-total correlations, which show whether each item contributes meaningfully to the composite; the conceptual coherence of the item set; and the width of the confidence interval around alpha, which reveals how much precision the sample size actually supports.

With n = 36, an alpha of 0.65 has an approximate 95% CI of [0.45, 0.80] under normal-theory assumptions. This wide interval reflects substantial sampling uncertainty: the true population reliability could range from questionable to good. Always report confidence intervals alongside point estimates. The default CI from `psych::alpha()` is based on an asymptotic standard error, which can be optimistic with small `n` or non-normal data; when computationally feasible, bootstrap intervals (for example via `boot::boot`) provide a useful robustness check.

⚠ Common Misconception: "Alpha > 0.70 = Good Scale"

Myth: "If Cronbach's alpha is above 0.70, my scale is reliable and valid."
Reality: Alpha can be **artificially inflated** by redundant items or scale length. High alpha ≠ good measurement.
Demonstration:

```
library(psych)

# Create a scale with near-duplicate items (bad scale design!)
set.seed(2025)
n <- 30
base_score <- rnorm(n, 50, 10)
redundant_scale <- tibble(
  item1 = base_score,
  item2 = base_score + rnorm(n, 0, 2),
  item3 = base_score + rnorm(n, 0, 2),
  item4 = base_score + rnorm(n, 0, 2)
)

# Compute alpha
```

```r
invisible(capture.output(
  alpha_result <- suppressWarnings(alpha(redundant_scale))
))

redundant_demo_display <- alpha_result$item.stats %>%
  as.data.frame() %>%
  rownames_to_column("Item") %>%
  transmute(
    Item,
    `Corrected item-total correlation` = round(r.cor, 2)
  )

chapter5_measurement_table(
  "5.1",
  "Corrected item-total correlations for the redundant-item
   ↪  demonstration",
  redundant_demo_display,
  note = paste0(
    "Cronbach's alpha = ",
    formatC(alpha_result$total$raw_alpha, format = "f", digits = 2),
    ". The high value reflects redundancy rather than broad construct
     ↪  coverage."
  ),
  align = c("l", "r")
)
```

Table 5.1

Corrected item-total correlations for the redundant-item demonstration

Item	Corrected item-total correlation
item1	0.99
item2	0.98
item3	0.98
item4	0.98

Note. Cronbach's alpha = 0.99. The high value reflects redundancy rather than broad construct coverage.
Problems with high-but-meaningless alpha:

1. **Redundant items**: Asking the same question 10 times inflates alpha but doesn't improve measurement
2. **Alpha increases with # items**: 20 poor items can yield $\alpha = 0.90$ (following the Spearman-Brown relationship)
3. **Ignores unidimensionality**: Alpha doesn't test whether items measure one construct or multiple

What to check instead: Ask whether all items contribute meaningfully to the scale through their item-total correlations, whether the mean inter-item correlation falls in a plausible range of about 0.15 to 0.50 (Briggs and Cheek 1986; Clark and Watson 1995), whether the items appear to reflect one coherent factor rather than several unrelated ones, and whether the content still covers the construct broadly enough to be useful. Values below about 0.15 suggest that items do not cohere well, whereas values above about 0.50 often indicate redundancy.
Lesson: **Don't chase alpha > 0.90 by adding redundant items.** Better to have $\alpha = 0.75$ with diverse, non-redundant items than $\alpha = 0.95$ with near-duplicates.

Example: Cronbach's Alpha for a Short Scale

We assess the internal consistency of a 3-item service quality scale using the `service_quality.csv` data.

```
library(tidyverse)
library(psych)

# Load service quality data
service_data <- read_csv("data/service_quality.csv", show_col_types =
 ↪  FALSE)

# Select the three quality items
quality_items <- service_data %>%
  select(q1_responsiveness, q2_professionalism, q3_clarity)

# Compute Cronbach's alpha
invisible(capture.output(
  alpha_result <- suppressWarnings(alpha(quality_items))
))

# Extract key values
alpha_est <- as.numeric(alpha_result$total$raw_alpha)
```

```r
alpha_se <- as.numeric(alpha_result$total$ase)

alpha_stat <- function(data, indices) {
  boot_items <- data[indices, , drop = FALSE]
  value <- suppressWarnings(psych::alpha(boot_items, warnings =
  ↪  FALSE)$total$raw_alpha)
  ifelse(is.finite(value), value, NA_real_)
}

set.seed(2025)
alpha_boot <- boot::boot(quality_items, statistic = alpha_stat, R = 1000)
alpha_boot_ci <- stats::quantile(alpha_boot$t[, 1], probs = c(0.025,
 ↪  0.975), na.rm = TRUE)

alpha_summary <- tibble(
  Metric = c(
    "Cronbach's alpha",
    "Asymptotic standard error",
    "Bootstrap 95% CI lower bound",
    "Bootstrap 95% CI upper bound",
    "Bootstrap resamples"
  ),
  Value = c(
    alpha_est,
    alpha_se,
    alpha_boot_ci[[1]],
    alpha_boot_ci[[2]],
    1000
  )
)

chapter5_measurement_table(
  "5.2",
  "Internal consistency summary for the 3-item service quality scale",
  alpha_summary,
  note = "The confidence interval is a percentile bootstrap interval over
   ↪  participants. The asymptotic standard error from psych::alpha() is
   ↪  shown for reference but is not used as the primary interval in this
   ↪  small-sample example.",
  align = c("l", "r"),
  digits = 3
)
```

Table 5.2

Internal consistency summary for the 3-item service quality scale

Metric	Value
Cronbach's alpha	0.793
Asymptotic standard error	0.060
Bootstrap 95% CI lower bound	0.613
Bootstrap 95% CI upper bound	0.881
Bootstrap resamples	1000.000

Note. The confidence interval is a percentile bootstrap interval over participants. The asymptotic standard error from psych::alpha() is shown for reference but is not used as the primary interval in this small-sample example.

Interpretation: Alpha quantifies the proportion of variance in scale scores attributable to the true score under the working assumptions of the model. Higher alpha indicates stronger internal consistency, but the bootstrap interval shows how uncertain the estimate remains in a small sample. If alpha is below 0.70, consider whether items truly measure a single construct or whether the scale is too heterogeneous. The `psych` package also reports "alpha if item deleted", showing how alpha would change if each item were removed; this helps identify problematic items.

Standard Error of Measurement (SEM)

The SEM quantifies measurement precision: how much individual scores vary because of measurement error. In applied reporting, the reliability value is usually an estimate such as $\hat{\alpha}$ rather than a known population parameter, so the SEM is also an estimate and should be interpreted with the same uncertainty caveat.

Formula: $\widehat{\text{SEM}} = \text{SD} \times \sqrt{(1 - \hat{\alpha})}$

```r
# Example: Test with SD = 10, estimated Cronbach's alpha = 0.75
scale_SD <- 10
alpha_hat <- 0.75

SEM <- scale_SD * sqrt(1 - alpha_hat)

# Confidence interval for individual scores
# 95% CI: observed score ± 1.96 × SEM
CI_width <- 1.96 * SEM

sem_summary <- tibble(
  Metric = c(
    "Scale SD",
    "Estimated alpha (alpha-hat)",
    "Estimated SEM",
    "95% CI half-width"
  ),
```

```
  Value = c(scale_SD, alpha_hat, SEM, CI_width)
)

chapter5_measurement_table(
  "5.3",
  "Standard error of measurement example for a 10-point test",
  sem_summary,
  align = c("l", "r"),
  digits = 2
)
```

Table 5.3

Standard error of measurement example for a 10-point test

Metric	Value
Scale SD	10.00
Estimated alpha (alpha-hat)	0.75
Estimated SEM	5.00
95% CI half-width	9.80

Interpretation: With $\widehat{\text{SEM}} = 5$ points, an individual's true score likely falls within about ±10 points of their observed score under the working reliability estimate. This helps judge whether observed changes are genuine or merely measurement error. In practice, the minimum detectable change is about $1.96 \times \widehat{\text{SEM}} \times \sqrt{2} \approx 14$ points, so smaller changes could plausibly reflect measurement error alone. Because $\hat{\alpha}$ is estimated from a small sample, this SEM should be reported as approximate; when individual-level decisions matter, use a confidence band for reliability, or a bootstrap sensitivity check, to show how much the SEM changes across plausible reliability values.

McDonald's Omega

McDonald's omega (ω_t) is an alternative to alpha that relaxes the tau-equivalence assumption (McDonald 1999). It is computed from a common-factor model and reflects the proportion of variance in scale scores due to common factors. In realistic congeneric settings, omega is often preferred over alpha because it accommodates unequal factor loadings across items (Trizano-Hermosilla and Alvarado 2016).

When to use: Multi-item scales with varying item-factor relationships, when tau-equivalence is questionable, or when reporting alongside alpha for robustness.

Example: McDonald's Omega

We compute omega for the same 3-item service quality scale.

```
library(psych)

# Compute McDonald's omega
invisible(capture.output(
  omega_result <- suppressWarnings(suppressMessages(omega(quality_items,
 ↪  nfactors = 1, plot = FALSE)))
))
invisible(capture.output(
  alpha_for_omega <- suppressWarnings(alpha(quality_items))
))

omega_tot <- as.numeric(omega_result$omega.tot)
alpha_for_omega_est <- as.numeric(alpha_for_omega$total$raw_alpha)

omega_summary <- tibble(
  Metric = c(
    "McDonald's omega total",
    "Cronbach's alpha",
    "Difference (omega - alpha)"
  ),
  Value = c(
    omega_tot,
    alpha_for_omega_est,
    omega_tot - alpha_for_omega_est
  )
)

chapter5_measurement_table(
  "5.4",
  "Omega summary for the 3-item service quality scale",
  omega_summary,
  note = "Omega and alpha are very similar here, which is what we expect
   ↪  for a tightly unidimensional toy example.",
  align = c("l", "r"),
  digits = 3
)
```

Table 5.4

Omega summary for the 3-item service quality scale

Metric	Value
McDonald's omega total	0.803

Metric	Value
Cronbach's alpha	0.793
Difference (omega - alpha)	0.010

Note. Omega and alpha are very similar here, which is what we expect for a tightly unidimensional toy example.

Interpretation: The `omega()` function reports two related indices. ω_t (omega total) summarises the proportion of total score variance attributable to all common factors, whereas ω_h (omega hierarchical) isolates the variance attributable to a general factor after accounting for group factors. For brief, unidimensional scales, ω_t is usually the relevant quantity. ω_h is more informative when evaluating multidimensional instruments with a bifactor-like structure.

With a perfectly unidimensional three-item example, some elements of the printed `omega()` output can look extreme. Values such as `omega_h = 1` or `max/min = Inf` reflect a degenerate single-factor solution with no modeled specific-factor variance. In this setting, they are a consequence of the toy example rather than evidence that the computation failed.

Split-Half Reliability

Split-half reliability divides a scale into two halves, computes the correlation between half-scale scores, and adjusts using the Spearman–Brown formula to estimate reliability of the full scale. The adjustment is $r_{SB} = \frac{2r_{half}}{1+r_{half}}$, where r_{half} is the correlation between the two halves. This matters because reliability increases with scale length. With small samples, split-half estimates are imprecise, and different item splits can yield noticeably different values.

When to use: Multi-item scales, desire for alternative reliability estimate, comparison with alpha or omega.

Example: Split-Half Reliability

We compute split-half reliability for the service quality scale and report the Spearman-Brown adjusted estimate.

```
library(psych)

set.seed(2025)

# Split-half reliability
invisible(capture.output(
  split_result <- splitHalf(quality_items, raw = TRUE)
))
```

```
split_summary <- tibble(
  Metric = c(
    "Spearman-Brown adjusted reliability",
    "Minimum split-half reliability",
    "Maximum split-half reliability",
    "Median split-half reliability"
  ),
  Value = c(
    as.numeric(split_result$alpha),
    as.numeric(split_result$minrb),
    as.numeric(split_result$maxrb),
    as.numeric(split_result$ci[["50%"]])
  )
)

chapter5_measurement_table(
  "5.5",
  "Split-half reliability summary for the service quality scale",
  split_summary,
  note = paste0(
    "Least favourable split: ",
    paste(split_result$minAB$A, collapse = ", "),
    " versus ",
    paste(split_result$minAB$B, collapse = ", "),
    "."
  ),
  align = c("l", "r"),
  digits = 3
)
```

Table 5.5

Split-half reliability summary for the service quality scale

Metric	Value
Spearman-Brown adjusted reliability	0.795
Minimum split-half reliability	0.670
Maximum split-half reliability	0.771
Median split-half reliability	0.679

Note. Least favourable split: q2_professionalism versus q1_responsiveness, q3_clarity.

Interpretation: The split-half correlation measures consistency between the two halves. The Spearman–Brown adjustment estimates the reliability of the full scale. This method is less commonly used than alpha but provides a complementary perspective. Because different item splits

can yield different results, it is best treated as a robustness check rather than a single definitive reliability estimate.

Worst Split-Half Reliability (Practical Alternative to Revelle's Beta)

This chapter reports the minimum split-half reliability across all admissible partitions from `psych::splitHalf()` as a practical stress test. Earlier editions of this text referenced Revelle's β, a related lower-bound estimate that focuses on the weakest split-half consistency; current `psych` workflows fold that logic into `splitHalf()`, so a separate helper is unnecessary. Unlike omega, which is factor-model based, the worst-split value is a stress test of how fragile the item set becomes under the least favourable partition.

```
invisible(capture.output(
  split_result <- splitHalf(quality_items, raw = TRUE)
))

worst_split_display <- tibble(
  Metric = c(
    "Least favourable split reliability",
    "Most favourable split reliability",
    "Least favourable split A",
    "Least favourable split B"
  ),
  Value = c(
    formatC(as.numeric(split_result$minrb), format = "f", digits = 3),
    formatC(as.numeric(split_result$maxrb), format = "f", digits = 3),
    paste(split_result$minAB$A, collapse = ", "),
    paste(split_result$minAB$B, collapse = ", ")
  )
)

chapter5_measurement_table(
  "5.6",
  "Worst-case split-half reliability for the service quality scale",
  worst_split_display,
  note = paste0(
    "This chapter reports the minimum split-half reliability from
     ↪  psych::splitHalf(). Earlier editions referenced Revelle's beta,
     ↪  but current psych workflows integrate that lower-bound logic
     ↪  within splitHalf() (here using psych ",
    as.character(packageVersion("psych")),
    ")."
  ),
  align = c("l", "l")
)
```

Table 5.6

Worst-case split-half reliability for the service quality scale

Metric	Value
Least favourable split reliability	0.670
Most favourable split reliability	0.771
Least favourable split A	q2_professionalism
Least favourable split B	q1_responsiveness, q3_clarity

Note. This chapter reports the minimum split-half reliability from psych::splitHalf(). Earlier editions referenced Revelle's beta, but current psych workflows integrate that lower-bound logic within splitHalf() (here using psych 2.6.3).

Interpretation: The minimum split-half value is typically lower than the Spearman-Brown adjusted estimate because it focuses on the weakest admissible partition. Large gaps between the overall split-half estimate and the minimum split suggest that some item partitions are fragile, meaning the scale could behave inconsistently across subsets of items. With very small samples, these stress-test values can fluctuate; report them alongside alpha and omega and acknowledge the additional uncertainty.

Polychoric Correlations for Ordinal Items

Likert-scale items (e.g., 1–7 ratings) are ordinal, not continuous. Pearson correlations and alpha computed on ordinal data may underestimate reliability. Polychoric correlations estimate the correlation between underlying continuous latent variables, assuming ordinal responses arise from categorising continuous variables.

When items are ordinal and have few response options, polychoric correlations and ordinal alpha may be more accurate. However, stable estimation often requires roughly n = 50–100, depending on the number of response categories and the observed distributions (Olsson 1979). With smaller samples, polychoric solutions may be unstable or fail to converge; in those cases, report Pearson-based alpha and note that ordinal methods were considered but were too uncertain for strong interpretation.

When to use: Ordinal items with few response categories, ideally with samples closer to n ≥ 50–100 when feasible, and when there is a clear need for theoretically appropriate latent-response correlations.

Example: Polychoric Correlations (Conceptual)

We compute polychoric correlations for the service quality items. With n = 36, this example falls below the preferred range for stable polychoric estimation, so the result should be treated as illustrative rather than definitive.

```
library(psych)

# Polychoric correlation matrix with error handling
# Falls back to Pearson correlations if sample is too small
poly_result <- tryCatch(
  polychoric(quality_items),
  error = function(e) {
    message("Polychoric estimation failed (sample too small). Using
     ↪  Pearson correlations.")
    return(list(rho = cor(quality_items)))
  }
)

# Compute alpha based on polychoric correlations
alpha_poly <- alpha(poly_result$rho)
alpha_poly_est <- as.numeric(alpha_poly$total$raw_alpha)

poly_display <- round(poly_result$rho, 3) %>%
  as.data.frame() %>%
  rownames_to_column("Item")

chapter5_measurement_table(
  "5.7",
  "Polychoric correlation matrix for the service quality items",
  poly_display,
  note = paste0(
    "Alpha based on the polychoric correlation matrix = ",
    formatC(alpha_poly_est, format = "f", digits = 3),
    "."
  ),
  align = c("l", "r", "r", "r"),
  digits = 3
)
```

Table 5.7

Polychoric correlation matrix for the service quality items

Item	q1_responsiveness	q2_professionalism	q3_clarity
q1_responsiveness	1.000	0.643	0.666
q2_professionalism	0.643	1.000	0.519

Item	q1_responsiveness	q2_professionalism	q3_clarity
q3_clarity	0.666	0.519	1.000

Note. Alpha based on the polychoric correlation matrix = 0.824.

Interpretation: Polychoric correlations are typically higher than Pearson correlations for ordinal data, so alpha computed from the polychoric matrix may also be higher. However, with small samples these estimates can be unstable or fail to converge. If polychoric and Pearson results are similar, the method choice has little practical impact. If they differ substantially, report both and note that the ordinal estimate is more assumption-sensitive in small samples.

Reporting Reliability with Small Samples

When reporting reliability for small samples and short scales:

- Report Cronbach's alpha with confidence intervals.
- Consider reporting McDonald's omega as a robustness check.
- Acknowledge limitations (short scale, small sample, wide CIs).
- Provide item-level descriptive statistics (means, SDs, inter-item correlations).
- Discuss implications for interpretation (e.g., "The modest alpha suggests caution in interpreting scale scores; findings should be replicated with longer instruments").

Lab Practical 5.1: Refining a Workplace Resilience Scale

Context: An organizational psychologist developed a 6-item Workplace Resilience Scale (WRS) to measure employees' ability to cope with job stress. After piloting the scale with 22 employees, the researcher wants to evaluate internal consistency and decide whether to drop any items to improve reliability. This walkthrough demonstrates item analysis, alpha calculation, and item-deletion decisions.

Learning Goals:

- Compute Cronbach's alpha for a multi-item scale
- Examine item-total correlations to identify weak items
- Assess the impact of dropping items on reliability
- Make evidence-based decisions about scale refinement
- Understand context-dependent alpha thresholds

Step 1: Load and Explore the Data

```
library(tidyverse)
library(psych)

# Simulated WRS data: 22 employees, 6 items (1-5 Likert scale)
set.seed(2025)
wrs_data <- tibble(
  WRS1 = c(4, 5, 3, 4, 5, 4, 3, 5, 4, 3, 4, 5, 3, 4, 4, 5, 3, 4, 5, 4, 3,
   ↪  4),
  WRS2 = c(3, 4, 3, 3, 4, 3, 2, 4, 3, 3, 4, 4, 3, 3, 4, 4, 3, 3, 4, 3, 2,
   ↪  3),
  WRS3 = c(5, 5, 4, 5, 5, 4, 4, 5, 5, 4, 5, 5, 4, 5, 5, 5, 4, 5, 5, 4, 4,
   ↪  5),
  WRS4 = c(2, 1, 3, 2, 1, 3, 4, 1, 2, 3, 2, 1, 3, 2, 1, 1, 3, 2, 1, 3, 4,
   ↪  2),  # Weak item
  WRS5 = c(4, 4, 3, 4, 4, 3, 3, 4, 4, 3, 4, 4, 3, 4, 4, 4, 3, 4, 4, 3, 3,
   ↪  4),
  WRS6 = c(3, 4, 3, 3, 4, 3, 2, 4, 3, 3, 4, 4, 3, 3, 4, 4, 3, 3, 4, 3, 2, 3)
)

# Descriptive statistics
wrs_descriptives <- wrs_data %>%
  summarise(across(everything(), list(mean = mean, sd = sd))) %>%
  pivot_longer(everything(), names_to = c("Item", ".value"), names_sep =
   ↪  "_") %>%
  rename(Mean = mean, SD = sd)

chapter5_measurement_table(
  "5.8",
  "Item descriptive statistics for the Workplace Resilience Scale pilot",
  wrs_descriptives,
  align = c("l", "r", "r"),
  digits = 2
)
```

Table 5.8

Item descriptive statistics for the Workplace Resilience Scale pilot

Item	Mean	SD
WRS1	4.00	0.76
WRS2	3.27	0.63
WRS3	4.64	0.49
WRS4	2.14	0.99
WRS5	3.64	0.49
WRS6	3.27	0.63

Checkpoint: Items have similar means (3–4) and SDs (0.5–1.0), except WRS4 shows a lower mean and higher SD. That pattern suggests it may not align with the rest of the scale and could be weakly keyed or reverse-worded.

Step 2: Compute Cronbach's Alpha (Full Scale)

```
# Compute alpha for all 6 items
invisible(capture.output(
  alpha_full <- suppressWarnings(alpha(wrs_data))
))

wrs_alpha_summary <- tibble(
  Metric = c(
    "Raw alpha",
    "Standardized alpha",
    "Average inter-item correlation",
    "Signal-to-noise ratio"
  ),
  Value = c(
    as.numeric(alpha_full$total$raw_alpha),
    as.numeric(alpha_full$total$std.alpha),
    as.numeric(alpha_full$total$average_r),
    as.numeric(alpha_full$total[["S/N"]])
  )
)

chapter5_measurement_table(
  "5.9",
  "Full-scale reliability summary for the 6-item WRS pilot",
  wrs_alpha_summary,
  note = "The negative WRS4 item pulls the full-scale estimate down sharply
   ↪  and should be inspected before any deletion decision is made.",
  align = c("l", "r"),
  digits = 3
)
```

Table 5.9

Full-scale reliability summary for the 6-item WRS pilot

Metric	Value
Raw alpha	0.101
Standardized alpha	0.625
Average inter-item correlation	0.218
Signal-to-noise ratio	1.669

Note. The negative WRS4 item pulls the full-scale estimate down sharply and should be inspected before any deletion decision is made.

Checkpoint: The raw alpha is very low here because one item is working in the opposite direction from the rest of the scale, not because every item is equally weak. The standardised alpha asks the same question after putting items on a common variance scale, the mean inter-item correlation shows how tightly the items move together, and the signal-to-noise ratio summarises how much reliable variance remains once error is considered. The key practical question is whether the "reliability if an item is dropped" table shows a sharp improvement for a specific item, because that pattern usually points to reverse coding or conceptual mismatch rather than to a uniformly poor scale.

Step 3: Examine Item-Total Correlations

```
# Extract item-total correlations
item_stats <- alpha_full$item.stats

# Display key statistics
item_stats_display <- item_stats %>%
  as.data.frame() %>%
  rownames_to_column("Item") %>%
  transmute(
    Item,
    `Corrected item-total correlation` = r.cor,
    `Raw item-total correlation` = r.drop,
    Mean = mean,
    SD = sd
  ) %>%
  arrange(`Corrected item-total correlation`)

chapter5_measurement_table(
  "5.10",
  "Item-total statistics for the 6-item WRS pilot",
  item_stats_display,
  align = c("l", "r", "r", "r", "r"),
  digits = 3
)
```

Table 5.10

Item-total statistics for the 6-item WRS pilot

Item	Corrected item-total correlation	Raw item-total correlation	Mean	SD
WRS4	-1.077	-0.981	2.136	0.990
WRS5	0.820	0.751	3.636	0.492
WRS3	0.820	0.751	4.636	0.492

Item	Corrected item-total correlation	Raw item-total correlation	Mean	SD
WRS2	0.857	0.805	3.273	0.631
WRS6	0.857	0.805	3.273	0.631
WRS1	0.943	0.809	4.000	0.756

Checkpoint: Look for:

- **r.cor**: Corrected item-total correlation (item vs. scale without that item). Values < 0.30 indicate weak contribution
- **r.drop**: Raw item-total correlation (item vs. full scale including that item)

Items with r.cor < 0.30 deserve close review. Here, WRS4 is negatively related to the rest of the scale, which points to either an unrecoded reverse-worded item or a construct mismatch rather than simple low reliability.

Step 4: Assess "Alpha if Item Dropped"

```
# Extract alpha-if-deleted
alpha_if_dropped <- alpha_full$alpha.drop

# Display with item labels
alpha_if_dropped_display <- alpha_if_dropped %>%
  as.data.frame() %>%
  rownames_to_column("Item") %>%
  transmute(
    Item,
    `Alpha if item dropped` = raw_alpha
  ) %>%
  arrange(desc(`Alpha if item dropped`))

chapter5_measurement_table(
  "5.11",
  "Alpha if each WRS item were removed",
  alpha_if_dropped_display,
  align = c("l", "r"),
  digits = 3
)
```

Table 5.11

Alpha if each WRS item were removed

Item	Alpha if item dropped
WRS4	0.937
WRS3	-0.504

Item	Alpha if item dropped
WRS5	-0.504
WRS2	-0.827
WRS6	-0.827
WRS1	-1.129

Checkpoint: Dropping WRS4 produces a very large improvement in alpha, which confirms that it is destabilizing the scale. In practice, the next step is to inspect the item wording and scoring key before deciding whether to delete it or reverse-score it.

Step 5: Recompute Alpha Without WRS4

```
# Drop WRS4 and recompute alpha
wrs_refined <- wrs_data %>% select(-WRS4)

invisible(capture.output(
  alpha_refined <- suppressWarnings(alpha(wrs_refined))
))

alpha_comparison <- tibble(
  Scale = c("Full 6-item scale", "Refined 5-item scale (WRS4 removed)"),
  Items = c("6", "5"),
  `Cronbach's alpha` = c(
    formatC(alpha_full$total$raw_alpha, format = "f", digits = 3),
    formatC(alpha_refined$total$raw_alpha, format = "f", digits = 3)
  )
)

chapter5_measurement_table(
  "5.12",
  "Reliability before and after removing WRS4",
  alpha_comparison,
  note = paste0(
    "Alpha increases by ",
    formatC(alpha_refined$total$raw_alpha - alpha_full$total$raw_alpha,
     ↪  format = "f", digits = 3),
    " after removing the misaligned item."
  ),
  align = c("l", "r", "r")
)
```

Table 5.12

Reliability before and after removing WRS4

Scale	Items	Cronbach's alpha
Full 6-item scale	6	0.101
Refined 5-item scale (WRS4 removed)	5	0.937

Note. Alpha increases by 0.836 after removing the misaligned item.

Checkpoint: The refined 5-item scale has much higher alpha and all remaining items show strong positive item-total correlations. That is a more internally consistent measure, but the jump is so large that it also suggests the remaining items may be quite redundant.

Step 6: Interpret in Context

Follow these guidelines:

- **Research/exploratory scales**: $\alpha \geq 0.60$–0.70 acceptable
- **Established scales in research**: $\alpha \geq 0.70$–0.80 preferred
- **High-stakes decisions (clinical, personnel)**: $\alpha \geq 0.80$–0.90 required
- **Short scales (3–5 items)**: Expect alpha to run about 0.05–0.10 lower than for longer scales with similar item quality

Decision: WRS4 requires immediate inspection because:

1. It has a negative corrected item-total correlation
2. Dropping it increases alpha dramatically
3. The remaining 5-item scale shows high internal consistency

If WRS4 was intentionally reverse-worded, recode it first and then rerun the reliability analysis before making a final deletion decision.

Step 7: Visualise Item Performance

Figure 5.1 shows the corrected item-total correlations for the pilot Workplace Resilience Scale. The dashed reference line marks the usual review threshold of about 0.30, making it easy to see why WRS4 demands immediate follow-up.

```
# Create item performance plot
item_performance <- item_stats %>%
  rownames_to_column("Item") %>%
  select(Item, r.cor, mean, sd)

ggplot(item_performance, aes(x = Item, y = r.cor, fill = r.cor > 0.30)) +
  geom_col() +
  geom_hline(yintercept = 0.30, linetype = "dashed", color = "red") +
  labs(
    title = "Item-Total Correlations (Corrected)",
    x = "Item",
    y = "Corrected Item-Total Correlation",
```

```
    fill = "Adequate (r > 0.30)"
  ) +
  theme_minimal() +
  scale_fill_manual(values = c("TRUE" = "steelblue", "FALSE" = "salmon"))
```

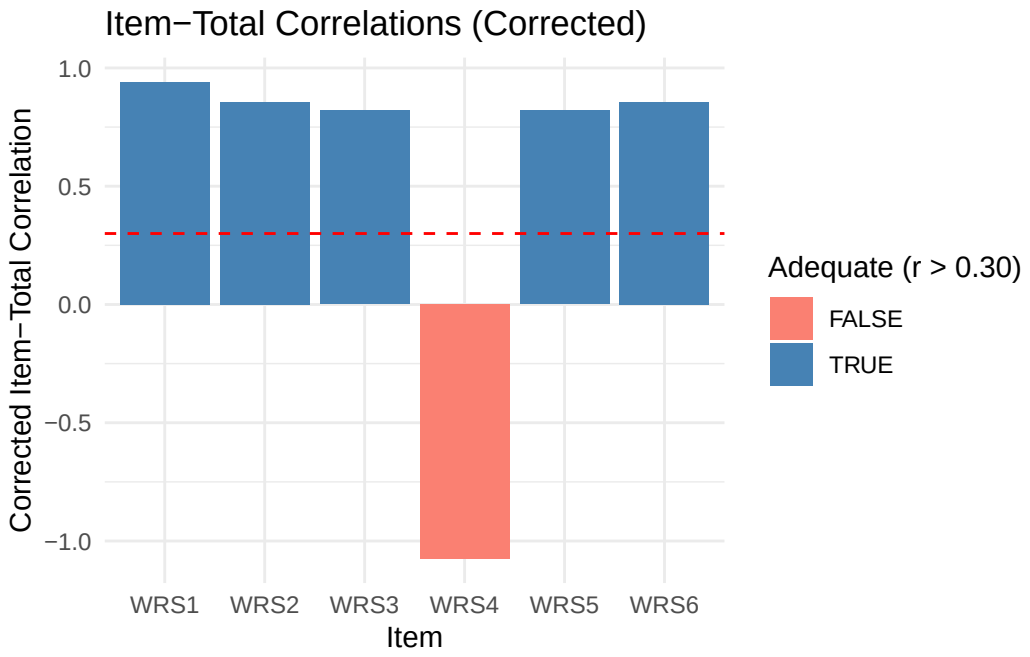


Figure 5.1: Corrected item-total correlations for the pilot Workplace Resilience Scale.

Checkpoint: The plot clearly shows that WRS4 falls well below the acceptable range and in fact moves in the opposite direction from the other items.

Step 8: Report the Results

> "The 6-item Workplace Resilience Scale was piloted with 22 employees. Initial reliability analysis suggested severe internal-consistency problems (raw alpha = 0.10), driven by WRS4, which showed a negative corrected item-total correlation (r = -0.98). Removing that item increased alpha to 0.94, indicating that the remaining five items were highly consistent. Before deleting WRS4 permanently, the researcher should verify whether it was intentionally reverse-worded and therefore requires reverse scoring rather than omission. Future validation should use a larger sample and include test-retest reliability."

Discussion Questions:

1. Why is item-total correlation more informative than item mean or SD? Item-total correlation shows how well an item captures the same construct as the rest of the scale. An item

can have a reasonable mean and standard deviation yet still dilute reliability if it does not move with the other items.

2. What if two items both have r.cor < 0.30? Review them one at a time rather than dropping both immediately. Recompute alpha after each change and check whether either item is reverse-worded, badly phrased, or intentionally measuring a distinct facet that the scale still needs.
3. Should you always maximise alpha by dropping items? No. Dropping items can improve internal consistency while also narrowing content coverage too far. In many applied settings, an alpha around 0.70 with broader construct coverage is preferable to a higher alpha achieved by keeping only a few overly redundant items.
4. How does sample size affect these decisions? Small samples make item-total correlations and alpha estimates unstable, so item decisions should remain tentative when n is very limited. Strong deletion decisions are best revisited once a larger validation sample is available.

Extension: Reverse-score WRS4 if its wording is intentionally opposite in direction, then compare the recoded solution with the deletion approach. You can also compute split-half reliability or omega to assess whether the remaining items behave as a coherent short scale.

Key Takeaways

Cronbach's alpha, McDonald's omega, and split-half estimates each provide useful but incomplete views of internal consistency. Short scales naturally produce lower coefficients than longer scales, and small samples make all of those estimates more uncertain, which is why confidence intervals and item-level diagnostics matter as much as the point estimate itself. Reliability should therefore be reported transparently, interpreted alongside content coverage and dimensionality, and treated as one part of measurement quality rather than a single threshold to clear.

Self-Assessment Quiz

Test your understanding of reliability and measurement quality from Chapter 5.

Question 1

Cronbach's alpha measures:

a) Whether data are normally distributed
b) Internal consistency—how closely related a set of items are
c) Test-retest reliability
d) Inter-rater agreement

Question 2

What is the main limitation of Cronbach's alpha?

a) It requires n>1,000
b) It assumes tau-equivalence (equal factor loadings across items)
c) It cannot be calculated for short scales
d) It is always too high

Question 3

A 3-item scale has α=0.55 with n=25. Is this acceptable?

a) No—alpha must always exceed 0.70
b) Possibly—short scales naturally have lower alpha; consider context, item-total correlations, and CI
c) Yes—always acceptable
d) No—the scale must be discarded

Question 4

McDonald's omega is preferred over alpha when:

a) Items have equal factor loadings
b) Items have varying factor loadings (not tau-equivalent)
c) Sample size exceeds 500
d) Data are categorical

Question 5

Item-total correlation measures:

a) How well an item correlates with the total scale score (excluding that item)
b) Test-retest stability
c) The mean of all items
d) Sample size adequacy

Question 6

What does “alpha if item deleted” show?

a) The p-value for each item
b) How alpha would change if a specific item were removed
c) The mean of each item
d) Whether items are normally distributed

Question 7

Polychoric correlations are used when:

a) Items are continuous and normally distributed
b) Items are ordinal (e.g., Likert scales) and you want to estimate correlations between underlying continuous latent variables
c) Sample size exceeds 1,000
d) Data have no missing values

Question 8

A scale has α=0.72 with n=36. The 95% CI is [0.52, 0.86]. What does this tell us?

a) Reliability is excellent
b) Reliability is moderate, but precision is limited (wide CI) due to small sample
c) The scale is unreliable
d) More items must be added

Question 9

Split-half reliability involves:

a) Testing participants twice
b) Dividing items into two halves, computing correlation, and applying Spearman-Brown correction
c) Removing half the sample
d) Using different raters

Question 10

A scale shows α=−0.15. What is the most likely cause?

a) Perfect reliability
b) Reverse-coded items not properly recoded, or items measuring different constructs
c) Sample size too large
d) Normal distribution

Answers and Explanations

Question 1

Answer: b)

Explanation: Cronbach's alpha quantifies internal consistency by comparing item variances to total scale variance. It indicates whether items measure a common construct. The chapter states: "Cronbach's alpha estimates internal consistency by comparing item variances to total scale variance."

Question 2

Answer: b)

Explanation: Alpha assumes all items contribute equally to the construct (tau-equivalence). When items have varying loadings, alpha may underestimate reliability. McDonald's omega relaxes this assumption. The chapter explicitly notes alpha "assumes that all items measure a single underlying construct with equal factor loadings (tau-equivalent model)."

Question 3

Answer: b)

Explanation: Alpha depends on scale length; 3-item scales often yield α=0.50-0.65 even when internally consistent. Check: (1) item-total correlations (all >0.30?), (2) alpha's 95% CI (precision), (3) conceptual coherence. For exploratory work, α=0.55 may be acceptable. This reflects the chapter's nuanced guidance about context-dependent thresholds rather than rigid cutoffs.

Question 4

Answer: b)

Explanation: Omega (ω_total) allows items to have different factor loadings, providing a more accurate reliability estimate when tau-equivalence does not hold. The chapter states: "McDonald's omega (ω) is an alternative to alpha that relaxes the tau-equivalence assumption."

Question 5

Answer: a)

Explanation: Corrected item-total correlation indicates how strongly each item relates to the overall scale. Values <0.30 suggest the item measures something different or is poorly worded. The lab practical emphasizes examining item-total correlations as a diagnostic tool.

Question 6

Answer: b)

Explanation: If alpha increases substantially when an item is removed, that item is weakening internal consistency (low correlation with others or measuring a different construct). Consider revising or removing it. This is a standard feature of reliability analysis output used to identify problematic items.

Question 7

Answer: b)

Explanation: Polychoric correlations assume ordinal responses arise from categorising underlying continuous variables. They often yield higher estimates than Pearson correlations for Likert data, but stable estimation often needs samples closer to n = 50–100. The chapter discusses them as theoretically appropriate for ordinal items, while also warning that very small samples can make them unstable.

Question 8

Answer: b)

Explanation: The point estimate (0.72) suggests acceptable reliability, but the wide CI reflects substantial uncertainty with n=36. The true population alpha could be as low as 0.52 (questionable) or as high as 0.86 (good). The chapter emphasizes that "small samples produce imprecise reliability estimates with wide confidence intervals."

Question 9

Answer: b)

Explanation: Split-half reliability divides a scale into two halves (e.g., odd vs even items), correlates the half-scores, then adjusts using the Spearman-Brown formula to estimate full-scale reliability. This is a classic alternative method for assessing reliability mentioned in the learning objectives.

Question 10

Answer: b)

Explanation: Negative alpha indicates that the average inter-item covariance is negative. This usually means reverse-coded items were not recoded, but it can also occur when the scale combines items tapping opposing constructs. Check the inter-item correlation matrix for negative values before interpreting the scale.

Chapter 6: Developing Short Scales for Small Samples

Learning Objectives

By the end of this chapter, you will be able to explain why short-scale development with small samples requires staged validation rather than one definitive study, match psychometric tools to the sample sizes they can realistically support, use item-level diagnostics to refine candidate items, and report scale-development evidence transparently without overstating what a pilot can show.

The Scale Development Lifecycle

Scale development is inherently a multi-stage process. With small samples, researchers must be strategic about which psychometric analyses to conduct at each stage and which to reserve for later validation.

The small helper functions used below (for example, `chapter6_simulate_scale()` and `chapter6_problem_item_diagnostics()`) are defined in `R/chapter_helpers.R` and sourced in the setup chunk. Readers who want to run individual chunks can first run `source("R/chapter_helpers.R")` from the project root.

The Iterative Process

Stage 1: Item Generation (n = 5–10 cognitive interviews)

A first round of about 5 to 10 cognitive interviews usually surfaces the clearest wording and comprehension problems in an initial item pool (Nielsen 1993).

Goal: Generate a pool of 2–3× your target number of items and ensure they are comprehensible.

Methods:

- **Literature review**: Identify existing scales and adapt items
- **Expert consultation**: Subject matter experts suggest relevant content

- **Cognitive interviews**: Think-aloud protocols with 5–10 participants from the target population

Example:

```
item_pool <- tibble(
  item_id = 1:15,
  item_text = c(
    "I feel confident handling work challenges",
    "I bounce back quickly from setbacks",
    "I maintain composure under pressure",
    "I remain calm when facing obstacles",
    "I trust my ability to solve problems",
    "I recover from stress effectively",
    "I rebound after difficult situations",
    "I regain focus after disruptions",
    "I process setbacks constructively",
    "I maintain perspective during challenges",
    "I adapt easily to changing priorities",
    "I adjust my approach when needed",
    "I seek support when needed",
    "I learn from difficult experiences",
    "I embrace new work demands"
  ),
  domain = rep(c("Confidence", "Recovery", "Adaptability"), each = 5),
  cognitive_issues = c(
    NA, "Ambiguous: what counts as 'quickly'?", NA, NA, NA,
    NA, "Too similar to item 2", NA, NA, NA,
    NA, NA, "Double-barreled: 'seek' and 'support'", NA, NA
  )
)

chapter6_scale_table(
  number = "6.1",
  title = "Candidate items flagged during cognitive interviewing.",
  data = item_pool %>%
    dplyr::filter(!is.na(cognitive_issues)) %>%
    dplyr::transmute(
      `Item ID` = item_id,
      `Candidate Item` = item_text,
      `Interview Note` = cognitive_issues
    ),
  note = "Items flagged during think-aloud work should be revised before
   ↪  any pilot administration.",
  align = c("r", "l", "l")
)
```

Table 6.1

Candidate items flagged during cognitive interviewing.

Item ID	Candidate Item	Interview Note
2	I bounce back quickly from setbacks	Ambiguous: what counts as 'quickly'?
7	I rebound after difficult situations	Too similar to item 2
13	I seek support when needed	Double-barreled: 'seek' and 'support'

Note. Items flagged during think-aloud work should be revised before any pilot administration.

Key Point: At this stage, **do NOT collect quantitative data**. Focus on qualitative feedback about item clarity, relevance, and comprehensiveness.

Stage 2: Pilot Testing (n = 20–30)

Goal: Identify problematic items before committing to a larger study.

Methods:

- Administer all items to a small pilot sample
- Compute **item-total correlations** (r.cor)
- Check for **ceiling/floor effects** (> 80% at extreme response)
- Examine **item means and SDs** (avoid items with no variance)

What you CAN do with n = 20–30:

At this stage, item-total correlations are screening tools rather than definitive item-deletion criteria. With `n = 20-30`, values near the usual `0.30` heuristic have substantial sampling variability, so a result like `0.28` versus `0.32` should be treated as a provisional flag for review rather than a mechanical keep-or-drop rule. The Briggs and Cheek (Briggs and Cheek 1986) mean-interitem-correlation guidance is useful here, but it is still a heuristic: for $n = 25$, a sample correlation of $r = 0.30$ has an approximate Fisher-z 95% CI from about -0.11 to 0.62. That interval is too wide to support automatic item deletion based on a single pilot correlation.

```
# Simulated pilot data: 25 participants, 12 items
pilot_data <- chapter6_simulate_scale(
  n = 25,
  loadings = c(0.90, 0.85, 0.80, 0.75, 0.70, 0.80, 0.78, 0.65, 0.82, 0.76,
  ↪  0.70, 0.74),
  noise_sd = 0.75,
  seed = 2025
)
pilot_data$WRS5 <- 5L
pilot_data$WRS8 <- sample(1:2, 25, replace = TRUE)
pilot_data$WRS11 <- sample(c(3L, 4L), 25, replace = TRUE, prob = c(0.15,
 ↪  0.85))
```

```r
pilot_diagnostics <- chapter6_problem_item_diagnostics(pilot_data)

chapter6_scale_table(
  number = "6.2",
  title = "Pilot-stage item diagnostics for the 12-item candidate scale.",
  data = pilot_diagnostics %>%
    dplyr::filter(Flag != "OK"),
  note = "The pilot flags one ceiling item, one weak item-total
   ↪  correlation, and one low-variance item for revision or removal.",
  align = c("l", "r", "r", "r", "r", "l")
)
```

Table 6.2

Pilot-stage item diagnostics for the 12-item candidate scale.

Item	Item-total r	Mean	SD	Extreme %	Flag
WRS5	NA	5.00	0.00	100	Ceiling/floor effect
WRS8	0.10	1.64	0.49	36	Low variance
WRS11	0.32	3.84	0.37	0	Low variance

Note. The pilot flags one ceiling item, one weak item-total correlation, and one low-variance item for revision or removal.

Interpretation:

- **WRS5**: Ceiling effect (100% of responses at the maximum). Remove or reword.
- **WRS8**: Weak item-total correlation of 0.10. Consider dropping or rewriting.
- **WRS11**: Low variance (`SD = 0.37`) and weak discrimination. Reword for clarity or replace it.

What you CANNOT do with n = 20–30:

> ⚠ Do NOT Overinterpret Alpha with n < 30
>
> Cronbach's alpha estimates are **highly unstable** with n < 30. The 95% confidence interval will usually be very wide. For example, when the observed alpha is around `0.65`, an approximate interval of about `[0.40, 0.85]` would not be unusual. The exact width depends on both the observed alpha value and the sample size, so use the `psych::alpha()` output to report the interval from your own data.
> **Instead**: Focus on item-level diagnostics (means, SDs, ceiling or floor patterns, and item-total correlations) to refine your scale. If software computes alpha while extracting those diagnostics, do not treat the pilot alpha as a stable reliability result. Defer formal reliability

reporting to Stage 3.

Stage 3: Refinement (n = 50–100)

Goal: Estimate reliability and assess dimensionality.

Methods:

- **Cronbach's alpha** with confidence intervals
- **McDonald's omega** (if you suspect multidimensionality)
- **Split-half reliability** as a robustness check
- **Exploratory Factor Analysis (EFA)** if n ≥ 100 and you suspect subscales

Example:

```
# Simulated refinement data: 60 participants, 8 items (problematic items
 ↪  removed)
refinement_data <- chapter6_simulate_scale(
  n = 60,
  loadings = c(0.85, 0.80, 0.78, 0.75, 0.82, 0.80, 0.77, 0.79),
  noise_sd = 0.95,
  seed = 2025
)

chapter6_scale_table(
  number = "6.3",
  title = "Refinement-stage reliability summary for the 8-item scale.",
  data = chapter6_reliability_summary(refinement_data),
  note = "The confidence interval is a percentile bootstrap interval from
   ↪  row resampling. Split-half values are Spearman-Brown-adjusted
   ↪  reliability coefficients from <code>psych::splitHalf()</code>.",
  align = c("l", "r")
)
```

Table 6.3

Refinement-stage reliability summary for the 8-item scale.

Metric	Value
Cronbach's alpha	0.833
Bootstrap 95% CI lower	0.747
Bootstrap 95% CI upper	0.885
Average inter-item correlation	0.383
Mean split-half reliability	0.832

Metric	Value
Minimum split-half reliability	0.779
Maximum split-half reliability	0.901

Note. The confidence interval is a percentile bootstrap interval from row resampling. Split-half values are Spearman-Brown-adjusted reliability coefficients from psych::splitHalf().

Interpretation:

- Here the scale shows good research reliability (`alpha = 0.835`) with an approximate 95% CI of about `[0.772, 0.898]`.
- The mean split-half reliability is `0.836`, with adjusted split-half values ranging from `0.789` to `0.895`, which supports the same conclusion from a second perspective.
- With `n = 60`, uncertainty is still present, but the interval is now narrow enough to support cautious reporting.

Exploratory Factor Analysis (EFA) with n = 50–100:

```
# EFA requires n  100 ideally; with n=60, results are exploratory only
# Simulate larger dataset for demonstration
efa_data <- chapter6_simulate_scale(
  n = 100,
  loadings = c(0.82, 0.79, 0.77, 0.74, 0.81, 0.78, 0.76, 0.80),
  noise_sd = 0.90,
  seed = 2027
)

# Parallel analysis is the key screening plot here; suppress console text
 ↪  in the rendered view
invisible(capture.output(
  fa.parallel(efa_data, fa = "fa")
))
```

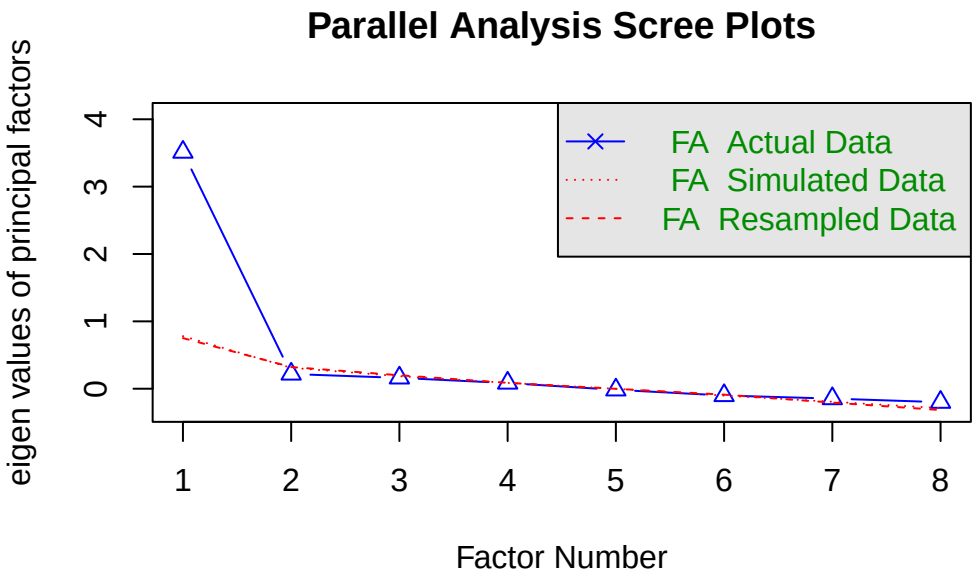


Figure 6.1: Parallel analysis for the candidate short scale.

```
# Fit 1-factor model
efa_result <- fa(efa_data, nfactors = 1, rotate = "oblimin", fm = "minres")
```

Figure 6.1 shows the parallel-analysis result for the simulated 8-item scale. In this example, the factor solution is the relevant guide: one factor is retained, and the fitted one-factor model explains about 43.9% of the total variance. Some software also prints a separate "components" recommendation. For early scale development, the factor recommendation is the substantively relevant result (Costello and Osborne 2005).

Caution: Even when parallel analysis suggests one dominant factor in a simulated example like this, `n = 100` still only supports **preliminary guidance**. Treat EFA results as provisional until a larger validation sample can confirm the structure. For early psychological scale development, a first factor explaining roughly 40% to 60% of the variance is often acceptable as an initial signal rather than a final validation result (Costello and Osborne 2005).

Stage 4: Validation (n = 150+)

Goal: Confirm scale structure and establish validity.

In this chapter, `SEM` refers to **structural equation modelling**. In Chapter 5, `SEM` denoted the **standard error of measurement**. The context distinguishes the two uses.

Methods:

- **Confirmatory Factor Analysis (CFA)**: Test hypothesized factor structure
- **Test-retest reliability**: Administer scale twice (2–4 weeks apart)

- **Convergent validity**: Correlate with theoretically related measures
- **Discriminant validity**: Show low correlation with unrelated constructs
- **Known-groups validity**: Scale discriminates between relevant groups

Example:

```
# CFA requires lavaan package and n  150
library(lavaan)

# Define 1-factor model
model <- '
  resilience =~ WRS1 + WRS2 + WRS3 + WRS4 + WRS5 + WRS6 + WRS7 + WRS8
'

# Fit model
cfa_result <- cfa(model, data = validation_data)
summary(cfa_result, fit.measures = TRUE, standardized = TRUE)

# Fit indices to assess model adequacy (conventional heuristics):
# - CFI >= 0.90 (acceptable), >= 0.95 (good)
# - RMSEA <= 0.08 (acceptable), <= 0.06 (good)
# - SRMR <= 0.08 (acceptable)
# Evaluate fit holistically rather than treating cutoffs as automatic
 ↪  rules.
```

Use these fit indices as heuristics rather than absolute pass-fail rules. CFI compares the hypothesised model with a baseline model in which the items are treated as unrelated. Larger values indicate better relative fit. RMSEA estimates lack of fit per model degree of freedom. Smaller values indicate closer approximate fit. SRMR summarises the average standardised residual discrepancy between the observed and model-implied correlations. Conventional thresholds such as CFI >= 0.95, RMSEA <= 0.06, and SRMR <= 0.08 are useful vocabulary, but they were developed mostly for larger samples and should not be applied mechanically in small validation studies (Hu and Bentler 1999). If fit is poor, report the indices, inspect the pattern of residuals and loadings, and explain the limitation. Do not repeatedly refit the model until the cutoffs are met.

Some short scales also contain both a broad general construct and narrower item clusters. In that setting, a bifactor model can separate the general factor from group factors and can support omega-hierarchical as an estimate of reliability for the general factor. This is a larger-sample validation tool, not a pilot-stage shortcut.

Test-retest reliability:

```
# Compute scale scores at Time 1 and Time 2
validation_data <- validation_data %>%
  dplyr::mutate(
```

```r
    resilience_t1 = rowMeans(dplyr::select(., WRS1_t1:WRS8_t1), na.rm =
     ↪  TRUE),
    resilience_t2 = rowMeans(dplyr::select(., WRS1_t2:WRS8_t2), na.rm =
     ↪  TRUE)
  )

# Intraclass correlation coefficient (ICC)
# Specify the model explicitly because ICC values depend on the chosen
 ↪  form.
library(irr)
icc_result <- icc(
  cbind(validation_data$resilience_t1, validation_data$resilience_t2),
  model = "twoway",
  type = "agreement",
  unit = "single"
)
print(icc_result)

# ICC > 0.75 is often treated as good test-retest reliability,
# but always report the ICC model and confidence interval.
```

When reporting ICC, specify the model and interpretation rule directly in the text. A two-way agreement ICC above about 0.75 is often treated as good evidence of test-retest reliability, but the exact value depends on the ICC form chosen (Cicchetti 1994; Koo and Li 2016).

Special Considerations for n $<$ 50

What You CANNOT Do

Analyses That Require Larger Samples

With n < 50, the following analyses are **not feasible** or will produce unreliable results:

1. **Exploratory Factor Analysis (EFA)**: Rule of thumb is n ≥ 100 or 5–10 participants per item
2. **Confirmatory Factor Analysis (CFA)**: Requires n ≥ 150–200 for stable parameter estimates
3. **Measurement Invariance Testing**: Requires n ≥ 200 per group
4. **Structural Equation Modelling (SEM)**: Complex models need n ≥ 200–400
5. **PLS-SEM (Partial Least Squares SEM)**: Despite "small-sample" marketing claims, stable path estimates still usually need at least n ≈ 100–150
6. **Item Response Theory (IRT)**: Most models require n ≥ 250–500
7. **Reliable Cronbach's alpha**: With n < 30, alpha estimates have 95% CIs spanning 0.3–0.4 units

Keep these methods for larger studies. Forced application to very small samples produces misleading results.

Why Structural Equation Modelling (SEM) Requires Large Samples

Question: "Can I use SEM, CFA, or PLS-SEM with my small sample (n < 100)?"
Short Answer: **No.** SEM-based methods require substantially larger samples than this book's target range (n = 10–100).

Minimum Sample Size Requirements

Method	Minimum n	Realistic n	Why?
Confirmatory Factor Analysis (CFA)	150	200-300	Stable factor loadings, fit indices
Structural Equation Modelling (SEM)	200	300-500	Complex path models, multiple latent variables
PLS-SEM	100	150-300	Despite "small-sample" marketing claims, stable path estimates usually need at least 100-150 cases
Multi-Group SEM	200/group	300/group	Measurement invariance testing

Rule of Thumb: 10-20 observations per estimated parameter is a common starting heuristic (e.g., 5 indicators + 3 paths = 8 parameters → need 80-160 observations), but actual requirements vary with model complexity, indicator quality, and estimation method.
That is why the table above should be read as planning guidance rather than a universal rulebook. Even methods sometimes marketed as "small-sample friendly," such as PLS-SEM, still need enough observations for stable path estimates and standard errors (Hair et al. 2017).

What Happens If You Ignore These Requirements?

With n < 100, SEM/CFA/PLS-SEM tends to produce unstable parameter estimates. Factor loadings can fluctuate sharply after very small data changes, path coefficients often carry very large standard errors, and conclusions may change when only a handful of observations are added or removed.
Small samples also increase the risk of non-convergent or improper solutions. Maximum-likelihood estimation may fail to converge, Heywood cases such as negative variances or loadings above 1.0 become more likely, and analysts can end up imposing arbitrary constraints simply to force the model to run.
Even when the software returns output, the fit statistics are difficult to trust. With small n, χ^2, CFI, TLI, and RMSEA can look reassuring for the wrong reasons, modification indices often suggest spurious changes, and a model may appear to fit the sample well only because it has overfit noise that will not replicate in new data.
The final danger is false confidence. SEM software will still print parameter estimates, p-values, confidence intervals, and polished path diagrams, but that appearance of technical completeness does not make the results trustworthy. Reviewers and readers will rightly question claims built on latent-variable models that the sample size cannot support.

What Should You Do Instead? (For n $<$ 100)

Use the methods in THIS book:

SEM Goal	Small-Sample Alternative	Chapter	Minimum n
Assess scale reliability	Cronbach's α, McDonald's ω, split-half	Ch 6	30-50
Validate items	Item-total correlations, alpha-if-deleted	Ch 6	30-50
Reduce dimensionality	**Sum/mean composite scores**	Ch 6	20+
Test relationships (X → Y)	Regression with composite scores	Ch 5	30-50

Multiple predictors	Penalized regression (ridge/lasso/elastic net)	Ch 13	50-100
Mediation (X → M → Y)	Simple mediation with bootstrap CIs	Part E Project 5	80-100
Latent correlations	Polychoric correlations (exploratory)	Ch 6	50+
Measurement precision	Standard Error of Measurement (SEM statistic)	Ch 6	30+

Key Principle: **Composite scores are your friend.** - Sum or average your scale items to create observed composite variables - Use these composites in regression, t-tests, ANOVA - Acknowledge measurement error in limitations section - Plan larger validation study (n ≥ 200) for future CFA/SEM

Example: Replacing SEM with Composite-Score Analysis

Proposed SEM Model (n = 60):

```
Job_Satisfaction (5 items) → Turnover_Intent (3 items)
     ↑
Performance (4 items)
```

Small-Sample Alternative:

```
#| context: interactive
# Create a small example dataset with 5 satisfaction, 4 performance,
# and 3 turnover-intention items.
set.seed(2025)
data <- tibble::as_tibble(
  matrix(
    sample(c(1:5, NA), size = 60 * 12, replace = TRUE, prob =
 ↪  c(rep(0.19, 5), 0.05)),
    nrow = 60
  )
)
names(data) <- c(
  paste0("satisfaction_", 1:5),
  paste0("performance_", 1:4),
  paste0("turnover_", 1:3)
)

# Create composite scores by averaging items when enough responses
 ↪  are present
data <- data %>%
  dplyr::mutate(
    satisfaction_missing = rowSums(is.na(dplyr::select(.,
 ↪  satisfaction_1:satisfaction_5))),
    performance_missing = rowSums(is.na(dplyr::select(.,
 ↪  performance_1:performance_4))),
    turnover_missing = rowSums(is.na(dplyr::select(.,
 ↪  turnover_1:turnover_3))),
    satisfaction = ifelse(
      satisfaction_missing <= 1,
      rowMeans(dplyr::select(., satisfaction_1:satisfaction_5), na.rm
 ↪  = TRUE),
```

```
      NA_real_
    ),
    performance = ifelse(
      performance_missing <= 1,
      rowMeans(dplyr::select(., performance_1:performance_4), na.rm =
↪  TRUE),
      NA_real_
    ),
    turnover = ifelse(
      turnover_missing == 0,
     rowMeans(dplyr::select(., turnover_1:turnover_3), na.rm = TRUE),
      NA_real_
    )
  )

# Test relationships with standard regression
model1 <- lm(turnover ~ satisfaction, data = data)
model2 <- lm(turnover ~ satisfaction + performance, data = data)

# Report coefficients, R², confidence intervals
```

Advantages: This composite-score approach works with n = 60, remains interpretable because coefficients refer to changes in averaged scale scores, and is usually more robust than forcing a latent-variable model onto limited data. It is also more honest, because it acknowledges that observed composites are being analysed directly rather than pretending the sample is large enough for stable latent-variable estimation.
Limitations to Acknowledge: Composite scores still contain measurement error, they cannot test complex factor structures, and they do not separate within-item from between-item variance. Missing-data rules also need to be stated explicitly so readers know when partial composites were allowed and when cases were excluded.
When to Pursue SEM: Collect n ≥ 200 in a follow-up study. Then: 1. Use EFA to explore factor structure (if theory is unclear) 2. Use CFA to confirm measurement model 3. Test structural paths with latent variables 4. Assess model fit rigorously

Software Will Let You Do Bad Things

Warning: SmartPLS, AMOS, Mplus, and other SEM software will happily run with n = 50. They will produce: - Parameter estimates - p-values - Fit indices - Pretty path diagrams **This does NOT mean the results are trustworthy.** Software cannot judge whether your sample size is adequate—**you must.**

Recommended Reading (For Future Large-Sample Studies)

When you collect n ≥ 200, consult these resources:

1. **Kline, R. B. (2016).** *Principles and Practice of Structural Equation Modeling* (4th ed.). Guilford Press.
    - Gold standard SEM textbook
    - Sample size guidelines (pp. 15-18, 264-270)
2. **Brown, T. A. (2015).** *Confirmatory Factor Analysis for Applied Research* (2nd ed.). Guilford Press.
    - CFA-specific guidance
    - Measurement invariance testing
3. **Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2022).** *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)* (3rd ed.). Sage.
    - PLS-SEM methods (but note: still needs n ≥ 100-150 realistically)
4. **Hoyle, R. H. (Ed.). (2023).** *Handbook of Structural Equation Modeling* (2nd ed.). Guilford Press.
    - Advanced topics, sample size planning

Bottom Line

For n = 10–100, the methods in this book are the defensible default because they are appropriate for small samples, robust to common assumption problems, honest about uncertainty, and interpretable for substantive readers and reviewers. Save SEM for a later study with n >= 200. Until then, composite scores combined with transparent regression-style analyses will usually serve the research question much better.

What You CAN Do

With n = 20–50, focus on these **feasible and informative** analyses:

1. **Content validity.** Use expert review, cognitive interviews, and explicit links to the theoretical framework to decide whether the item pool is clear, relevant, and broad enough before you rely on any statistics.
2. **Item-level diagnostics.** Examine item means, standard deviations, skewness, corrected item-total correlations, floor or ceiling effects, and the inter-item correlation matrix to flag items that are obviously misbehaving.

3. **Preliminary reliability (n ≥ 30).** Report Cronbach's alpha with its 95% confidence interval, add split-half reliability as a robustness check, and inspect the mean inter-item correlation to see whether the scale is too loose or too redundant.
4. **Known-groups validity.** Compare scale scores between groups that theory predicts should differ, and use nonparametric methods such as the Mann–Whitney U test if the scale scores are skewed or ordinal.
5. **Preparation for a larger validation study.** Document the item-generation process, report pilot decisions transparently including which items were dropped and why, and use the pilot to specify the hypotheses that a later CFA or validation study will test.

Example: Documenting a Small-Sample Scale Development

```
# Reporting template for n < 50 pilot study
scale_dev_report <- tibble(
  Stage = c("Item Generation", "Pilot Testing", "Refinement",
   ↪  "Validation"),
  Sample_Size = c("n = 8 (cognitive interviews)", "n = 25", "n = 60",
   ↪  "Planned: n = 200"),
  Analyses_Conducted = c(
    "Think-aloud protocols; expert review (CVI = 0.88)",
    "Item-total correlations; ceiling/floor checks; 3 items dropped",
    "Alpha = 0.84 [95% CI: 0.77, 0.90]; split-half = 0.79-0.90",
    "CFA, test-retest ICC, convergent validity (planned)"
  ),
  Key_Findings = c(
    "Generated 15 items; 2 flagged as ambiguous",
    "Dropped WRS5 (ceiling), WRS8 (r.cor = 0.10), WRS11 (low variance)",
    "8-item scale shows acceptable reliability for research use",
    "Pending larger validation sample"
  )
)

chapter6_scale_table(
  number = "6.4",
  title = "Illustrative staged reporting summary for a short-scale
   ↪  development project.",
  data = scale_dev_report %>%
    dplyr::rename(
      `Development Stage` = Stage,
      `Sample Size` = Sample_Size,
      `Analyses Conducted` = Analyses_Conducted,
      `Key Findings` = Key_Findings
    ),
  note = "Report the actual pilot sample sizes, item deletions, and
   ↪  planned validation targets transparently.",
```

```
  align = c("l", "l", "l", "l")
)
```

Table 6.4

Illustrative staged reporting summary for a short-scale development project.

Development Stage	Sample Size	Analyses Conducted	Key Findings
Item Generation	n = 8 (cognitive interviews)	Think-aloud protocols; expert review (CVI = 0.88)	Generated 15 items; 2 flagged as ambiguous
Pilot Testing	n = 25	Item-total correlations; ceiling/floor checks; 3 items dropped	Dropped WRS5 (ceiling), WRS8 (r.cor = 0.10), WRS11 (low variance)
Refinement	n = 60	Alpha = 0.84 [95% CI: 0.77, 0.90]; split-half = 0.79-0.90	8-item scale shows acceptable reliability for research use
Validation	Planned: n = 200	CFA, test-retest ICC, convergent validity (planned)	Pending larger validation sample

Note. Report the actual pilot sample sizes, item deletions, and planned validation targets transparently.

Reporting Guidelines for Small-Sample Scale Development

When publishing or reporting scale development work with n < 50:

1. **Acknowledge limitations explicitly**:
    - "With n = 25, we could not reliably estimate Cronbach's alpha. Instead, we focused on item-total correlations to identify weak items."
    - "Exploratory factor analysis was not feasible (n = 60); we plan CFA with a larger sample (target n = 200)."
2. **Report what you did (not what you wish you could do)**: Avoid presenting alpha as a definitive reliability estimate when n < 30. If a journal requires it, report the confidence interval and describe the result as preliminary screening information. Do not force EFA or CFA with inadequate samples. Instead, explain the theoretical rationale for the proposed item grouping and reserve the structural test for a later study.

3. **Frame as a preliminary/pilot study**: State plainly that the work is a pilot or refinement study and explain what that means for interpretation. For example: "This pilot study (n = 35) was designed to refine item wording and identify problematic items before a larger validation study," or "Results are preliminary and should be interpreted with caution pending validation with n ≥ 150."

4. **Provide detailed item-level information**:

    - Publish item means, SDs, item-total correlations
    - Report which items were dropped and why
    - Share cognitive interview feedback (qualitative)

5. **Plan and fund the validation study**:

    - Use pilot data to justify sample size for validation (power analysis for CFA)
    - Secure resources for n ≥ 150–200 before claiming a "validated" scale

Key Takeaways

Scale development with small samples is best understood as a staged process rather than a single psychometric event. Very small pilots are most useful for qualitative refinement and item-level diagnostics, somewhat larger samples can support cautious preliminary reliability work, and only much larger studies can justify structural validation through EFA, CFA, and broader validity testing. Across all of those stages, transparency about what was and was not feasible is what makes the evidence credible.

Self-Assessment Quiz

Question 1

Why is scale development with small samples best treated as a multi-stage process?

a) Because all psychometric evidence can be collected from one pilot sample
b) Because different stages support different goals, from item clarity to later validation, and each requires different sample sizes
c) Because short scales never need validation
d) Because only qualitative methods are allowed with small samples

Question 2

At the item-generation stage, what is the main priority?

a) Estimating Cronbach's alpha
b) Running confirmatory factor analysis
c) Checking whether items are clear, relevant, and comprehensible through literature review, expert consultation, and cognitive interviews
d) Computing test-retest reliability

Question 3

With a pilot sample of about 20 to 30 participants, which analysis is most appropriate?

a) A full CFA with fit indices
b) Item-total correlations, ceiling or floor checks, and item means and SDs
c) A definitive test-retest reliability study
d) A published claim that the scale is fully validated

Question 4

Why does the chapter warn against relying on Cronbach's alpha with n < 30?

a) Alpha cannot be computed in R
b) Alpha estimates are highly unstable and have very wide confidence intervals with such small samples
c) Alpha is only for binary items
d) Alpha is irrelevant for short scales

Question 5

What should you do when an item shows a ceiling effect in the pilot data?

a) Automatically keep it because everyone agrees
b) Remove or reword it because respondents are not differentiating on that item
c) Use CFA immediately
d) Average it with another item

Question 6

At roughly what stage does the chapter consider exploratory factor analysis potentially feasible?

a) Immediately after cognitive interviews
b) At the refinement stage, ideally when n approaches or exceeds 100, and even then results are tentative with modest samples
c) Only after test-retest reliability
d) Never in small-sample research

Question 7

What is the main goal of the validation stage?

a) To decide whether items are comprehensible
b) To confirm scale structure and establish validity evidence such as CFA, test-retest reliability, and convergent validity
c) To compute item means only
d) To avoid collecting any more data

Question 8

Which reporting practice best fits a pilot scale-development study with n = 35?

a) Claim the scale is fully validated
b) Report preliminary item-level diagnostics, acknowledge the limitations, and state that larger-sample validation is still needed
c) Suppress the sample size because it looks weak
d) Present CFA fit indices from an underpowered model as definitive evidence

Question 9

Why is transparency especially important in small-sample scale development?

a) Because reviewers ignore sample sizes
b) Because readers need to know which psychometric claims are supported now and which must wait for later validation
c) Because short scales do not need theory
d) Because only qualitative evidence matters

Question 10

Which conclusion is most consistent with the chapter's overall message?

a) A small pilot sample can fully validate a new instrument
b) Scale development with small samples can be rigorous if it proceeds in stages and reserves stronger psychometric claims for later, larger studies
c) Factor analysis is always mandatory before any scale can be used
d) Item wording matters less than alpha

Answers and Explanations

Question 1

Answer: b)

Explanation: The chapter presents scale development as an iterative process. Early stages focus on item generation and pilot diagnostics, while later stages support reliability estimation, factor analysis, and validation. Small samples can support some of these steps, but not all of them at once.

Question 2

Answer: c)

Explanation: Stage 1 emphasises item generation and comprehension. The chapter explicitly states that this stage should focus on literature review, expert consultation, and cognitive interviews rather than quantitative psychometric estimation.

Question 3

Answer: b)

Explanation: The chapter recommends item-level diagnostics at the pilot stage: item-total correlations, ceiling and floor checks, and basic descriptive statistics. These analyses help identify weak items before larger validation work.

Question 4

Answer: b)

Explanation: The warning callout explains that alpha estimates are highly unstable with n < 30, making the confidence interval so wide that the estimate is not very informative for decision-making.

Question 5

Answer: b)

Explanation: The chapter's pilot example flags ceiling effects as a sign that an item may not discriminate well. The recommended response is to remove or reword the item.

Question 6

Answer: b)

Explanation: The chapter states that EFA is more defensible at the refinement stage and ideally with n around 100 or more. With smaller samples, the results are described as exploratory and unstable.

Question 7

Answer: b)

Explanation: Stage 4 is the validation phase. It is where the chapter places confirmatory factor analysis, test-retest reliability, convergent validity, discriminant validity, and known-groups validity.

Question 8

Answer: b)

Explanation: The reporting guidelines emphasise honesty about what the study could and could not show. With a pilot sample, the chapter recommends framing findings as preliminary and stating the need for a larger validation study.

Question 9

Answer: b)

Explanation: The chapter repeatedly stresses transparency: report what you did, what you did not do, and why. Small samples can support useful refinement decisions, but they do not justify overclaiming full validation.

Question 10

Answer: b)

Explanation: Small samples require staged, realistic claims. Early work can be rigorous when researchers match their analyses to the evidence their sample can support, even if full validation must wait for a larger dataset.

Chapter 7: Data Screening and Diagnostic Checks

Learning Objectives

By the end of this chapter, you will be able to explain why small samples are especially sensitive to outliers and assumption violations, detect unusual or influential observations with standard R diagnostics, distinguish plausible extreme values from likely data errors, and document defensible screening decisions transparently before formal analysis.

Why Data Screening Matters More with Small Samples

A single outlier can dominate a mean, distort a correlation, or violate regression assumptions when samples are small. Data entry errors (typos, misplaced decimals, incorrect codes) are harder to detect with fewer observations. Distributional assumptions (normality, homoscedasticity) are harder to verify with small samples, yet violations have greater consequences.

Systematic data screening before analysis helps identify problems early. Document all cleaning and transformation decisions in a reproducible script. Report descriptive statistics, missingness patterns, and any deviations from planned analyses.

A Practical Screening Workflow

Data screening should follow a fixed order so that decisions are not driven by the result of the main hypothesis test. Start with structural checks: confirm that IDs are unique, variables use the intended coding scheme, impossible values are absent, and missing values are represented consistently. Next, inspect univariate distributions with summaries and plots. Then inspect relationships among variables, including scatterplots for continuous variables and cross-tabulations for categorical variables. Only after those checks should model-specific diagnostics such as residual plots, leverage, Cook's distance, VIFs, and robust standard-error sensitivity checks be used.

The workflow below is a useful default for small-sample projects.

Step	Question	Typical check	Action if flagged
1	Are values structurally valid?	Ranges, labels, duplicate IDs	Correct from source records or document exclusion
2	Are missing values patterned?	Missingness table or plot	Describe pattern; decide whether complete-case analysis is defensible
3	Are any observations unusual on one variable?	Dotplots, boxplots, z-scores, Tukey fences	Inspect source record; retain genuine values unless outside target population
4	Are observations unusual jointly?	Scatterplots, Mahalanobis distance	Treat as candidates for inspection, not automatic exclusions
5	Are regression assumptions visibly strained?	Residual plots, Q-Q plots, leverage, Cook's distance	Report sensitivity analyses; consider robust or nonparametric alternatives
6	Are predictors redundant?	Correlations, VIFs, condition number	Simplify the model, combine variables, or avoid interpreting individual slopes

This order matters. For example, a high Cook's distance is less informative if the value is a data-entry error; a normality test is less useful if the outcome is visibly ordinal; and a regression coefficient is hard to interpret if two predictors measure almost the same construct. The screening report should therefore describe the sequence of checks, not just list isolated diagnostics.

Detecting Outliers

Outliers are observations that are unusually large or small relative to the rest of the data. They may represent legitimate extreme values, data entry errors, or individuals from a different population. With small samples, outliers can have disproportionate influence on results. A sensible screening sequence combines visual inspection through boxplots, histograms, or scatterplots with numerical rules such as Tukey's fences or extreme z-scores, and then moves to influence diagnostics such as leverage and Cook's distance when a regression model is involved. Outliers should be removed only when there is clear evidence of data entry error or that the observation does not belong to the target population. Otherwise, the better practice is to document the case, analyse the data with and without it if necessary, and explain how it affects the results.

Multivariate Outliers with Mahalanobis Distance

Univariate rules may miss cases that are unusual only when variables are considered jointly. Mahalanobis distance measures how far an observation lies from the multivariate centre, accounting for covariances among variables. Distances can be compared to the 97.5th percentile of the chi-square distribution with df equal to the number of variables screened jointly (Huberty and Olejnik 2006). This corresponds to a two-sided 5% significance level for flagging multivariate outliers. With small samples, covariance estimates are unstable; treat flagged cases as candidates for inspection, not automatic exclusion.

```
library(tidyverse)
library(MASS)

set.seed(2025)

# Simulated bivariate data: negative correlation between satisfaction and
 ↪  wait time
# Outlier is elevated on BOTH - unusual only in the multivariate sense
mu    <- c(satisfaction = 5.5, wait_time = 12)
Sigma <- matrix(c(1, -1.5, -1.5, 4), nrow = 2,
                dimnames = list(c("satisfaction", "wait_time"),
                 ↪  c("satisfaction", "wait_time")))
core  <- as.data.frame(mvrnorm(18, mu = mu, Sigma = Sigma))
core$satisfaction <- round(pmin(pmax(core$satisfaction, 1), 7), 1)
core$wait_time    <- round(pmax(core$wait_time, 1), 1)
multi_data <- bind_rows(core, tibble(satisfaction = 6.8, wait_time =
 ↪  18.5))  # multivariate outlier

center <- colMeans(multi_data)
cov_mat <- cov(multi_data)

multi_data <- multi_data %>%
  mutate(
    mahal = mahalanobis(., center, cov_mat),
    flag = mahal > qchisq(0.975, df = ncol(multi_data))
  )

multi_display <- multi_data %>%
  mutate(
    satisfaction = round(satisfaction, 2),
    wait_time = round(wait_time, 2),
    mahal = round(mahal, 2),
    flag = if_else(flag, "Yes", "No")
  ) %>%
  rename(
    Satisfaction = satisfaction,
    `Wait time` = wait_time,
    `Mahalanobis distance` = mahal,
```

```
    Flagged = flag
  )

smallsamplelab_apa_table(
  "7.1",
  "Mahalanobis distance results for the simulated bivariate dataset",
  multi_display,
  note = "Cases flagged as Yes exceed the 97.5th percentile of the
   ↪  chi-square distribution for two variables.",
  align = c("r", "r", "r", "c")
)
```

Table 7.1

Mahalanobis distance results for the simulated bivariate dataset

Satisfaction	Wait time	Mahalanobis distance	Flagged
5.7	13.5	0.31	No
4.1	11.5	2.17	No
5.0	13.6	0.38	No
4.0	14.4	2.75	No
3.6	12.1	3.56	No
5.3	11.5	0.33	No
5.2	12.8	0.04	No
6.3	12.1	0.96	No
6.3	11.5	1.18	No
4.9	13.4	0.39	No
6.0	11.3	0.82	No
6.5	8.3	5.86	No
6.4	11.4	1.44	No
3.8	13.1	2.74	No
5.3	14.4	0.79	No
5.3	11.5	0.33	No
5.9	12.0	0.37	No
5.9	13.1	0.35	No
6.8	18.5	11.21	Yes

Note. Cases flagged as Yes exceed the 97.5th percentile of the chi-square distribution for two variables.

Interpretation: Table 7.1 shows which observations exceed the Mahalanobis cutoff. Flagged cases should then be inspected individually to determine whether they reflect data errors, rare but valid combinations, or participants from a different subpopulation. Because covariance estimates are

noisy in small samples, report the cutoff explicitly rather than implying that the threshold is universal.

Example: Outlier Detection with Boxplots and Z-Scores

We examine a small dataset of customer wait times (n = 20) and check for outliers. Figure 7.1 shows the distribution visually, while Table 7.2 records which case is flagged by the IQR and z-score rules.

```
library(tidyverse)

set.seed(2025)

# Simulated wait times (most between 5-15 minutes, one outlier at 45)
wait_times <- c(7, 9, 8, 11, 10, 12, 8, 9, 10, 11, 13, 9, 10, 12, 8, 11,
 ↪  10, 9, 45, 10)

# Create the data frame
wait_data <- tibble(observation = 1:20, wait_time = wait_times)

# Boxplot
print(
  ggplot(wait_data, aes(x = "", y = wait_time)) +
    geom_boxplot(fill = "lightblue") +
    labs(title = "Boxplot of Wait Times", x = NULL, y = "Wait Time
     ↪  (minutes)") +
    theme_minimal() +
    theme(
      axis.text.x = element_blank(),
      axis.ticks.x = element_blank(),
      panel.grid.major.x = element_blank(),
      panel.grid.minor.x = element_blank()
    )
)
```

Figure 7.1: Customer wait-time boxplot with a potential outlier.

```r
# Identify outliers using 1.5*IQR rule
Q1 <- quantile(wait_times, 0.25)
Q3 <- quantile(wait_times, 0.75)
IQR_val <- IQR(wait_times)
lower_fence <- Q1 - 1.5 * IQR_val
upper_fence <- Q3 + 1.5 * IQR_val

# Z-scores for outlier identification
wait_data <- wait_data %>%
  dplyr::mutate(z_score = (wait_time - mean(wait_time)) / sd(wait_time))

flagged_cases <- wait_data %>%
  dplyr::mutate(
    `Z score` = round(z_score, 2),
    `IQR flag` = if_else(wait_time < lower_fence | wait_time >
     ↪  upper_fence, "Yes", "No"),
    `|z| > 3` = if_else(abs(z_score) > 3, "Yes", "No")
  ) %>%
  dplyr::filter(`IQR flag` == "Yes" | `|z| > 3` == "Yes") %>%
  dplyr::transmute(
    Observation = observation,
    `Wait time` = wait_time,
    `Z score`,
    `IQR flag`,
    `|z| > 3`
  )
```

```
smallsamplelab_apa_table(
  "7.2",
  "Outlier flags for the customer wait-time example",
  flagged_cases,
  note = "The same case is flagged by both the 1.5 × IQR rule and the
   ↪  absolute z-score criterion greater than 3.",
  align = c("r", "r", "r", "c", "c")
)
```

Table 7.2

Outlier flags for the customer wait-time example

Observation	Wait time	Z score	IQR flag	|z| > 3
19	45	4.17	Yes	Yes

Note. The same case is flagged by both the 1.5 × IQR rule and the absolute z-score criterion greater than 3.

Figure 7.1: Customer wait-time boxplot with a potential outlier.

Interpretation: Figure 7.1 visually flags observation 19 (wait time = 45 minutes), and Table 7.2 confirms that the same case is identified by both the IQR-based rule and the z-score criterion. Before removing it, investigate whether it reflects a data entry error or a genuine but unusual event. If it is genuine, a defensible alternative is to report results with and without the case or use robust methods that are less sensitive to extremes.

Checking Normality

Many parametric tests assume normally distributed data (or residuals). With small samples, normality is hard to verify formally. Visual checks (histograms, Q-Q plots) are more informative than statistical tests (Shapiro-Wilk), which have low power with small n. A non-significant Shapiro-Wilk result ($p > 0.05$) with $n < 30$ does not confirm normality; it may simply reflect insufficient power to detect departures. Conversely, a significant result warrants visual inspection to judge practical severity.

If data are clearly skewed or have heavy tails, consider nonparametric methods, transformations such as the log or square root, or robust procedures such as trimmed means and bootstrap intervals. This step is about matching the analysis to the actual shape of the data rather than conforming the data to parametric assumptions.

Example: Q-Q Plot for Normality Assessment

We check whether a small sample of test scores (n = 18) is approximately normally distributed. Figure 7.2 provides the visual diagnostic, and Table 7.3 reports the Shapiro-Wilk statistics in the same rounding format used in the text.

```
library(tidyverse)

set.seed(2025)

# Simulated test scores (approximately normal)
test_scores <- round(rnorm(18, mean = 70, sd = 10))

# Q-Q plot
qqnorm(test_scores, main = "Q-Q Plot of Test Scores")
qqline(test_scores, col = "red")
```

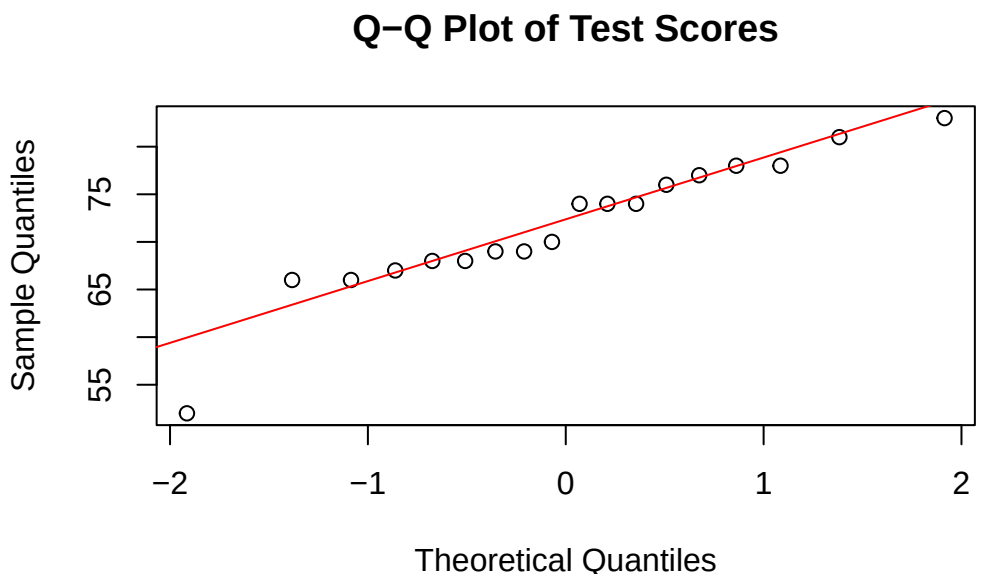


Figure 7.2: Q-Q plot of test scores.

```
# Shapiro-Wilk test
shapiro_result <- shapiro.test(test_scores)

shapiro_display <- tibble(
  Test = "Shapiro-Wilk",
  W = formatC(unname(shapiro_result$statistic), format = "f", digits = 4),
  `p value` = formatC(shapiro_result$p.value, format = "f", digits = 4)
)
```

```r
smallsamplelab_apa_table(
  "7.3",
  "Normality test results for the simulated test-score sample",
  shapiro_display,
  note = "The p-value is shown to four decimals so the table matches the
   ↪  printed test output exactly.",
  align = c("l", "r", "r")
)
```

Table 7.3

Normality test results for the simulated test-score sample

Test	W	p value
Shapiro-Wilk	0.9231	0.1465

Note. The p-value is shown to four decimals so the table matches the printed test output exactly.

Figure 7.2: Q-Q plot of test scores.

Interpretation: Figure 7.2 shows whether the points remain close to the diagonal line expected under normality, while Table 7.3 reports the Shapiro-Wilk result in matching precision. Here the p-value does not provide strong evidence against normality, but it does not prove normality. With a small sample, the visual diagnostic remains more informative than the test alone.

Linearity and Homoscedasticity in Regression

Linear regression assumes a linear relationship between predictors and outcome, and constant variance of residuals (homoscedasticity). Scatterplots of residuals vs. fitted values help assess these assumptions.

In a residuals-versus-fitted plot, linearity is supported when residuals scatter randomly around zero with no systematic curvature, and homoscedasticity is supported when the vertical spread remains roughly constant. With n about 20, these patterns are inherently noisy; focus on gross violations such as clear funnel shapes or strong curvature rather than subtle deviations.

Example: Regression Diagnostics

We fit a simple linear regression (outcome ~ predictor) with n = 20 and check the standard diagnostic panel. Figure 7.3 summarises the model diagnostics, and Table 7.4 reports the fitted coefficients so the regression example has a formatted numerical counterpart rather than a console dump.

```
library(tidyverse)

set.seed(2025)

# Simulated study-hours / exam-score data (n = 20)
reg_data <- tibble(
  study_hours = round(runif(20, 1, 12)),
  exam_score  = 40 + 4.5 * study_hours + rnorm(20, 0, 8)
)

# Fit linear model
model <- lm(exam_score ~ study_hours, data = reg_data)

# Diagnostic plots
par(mfrow = c(2, 2))
plot(model)
```

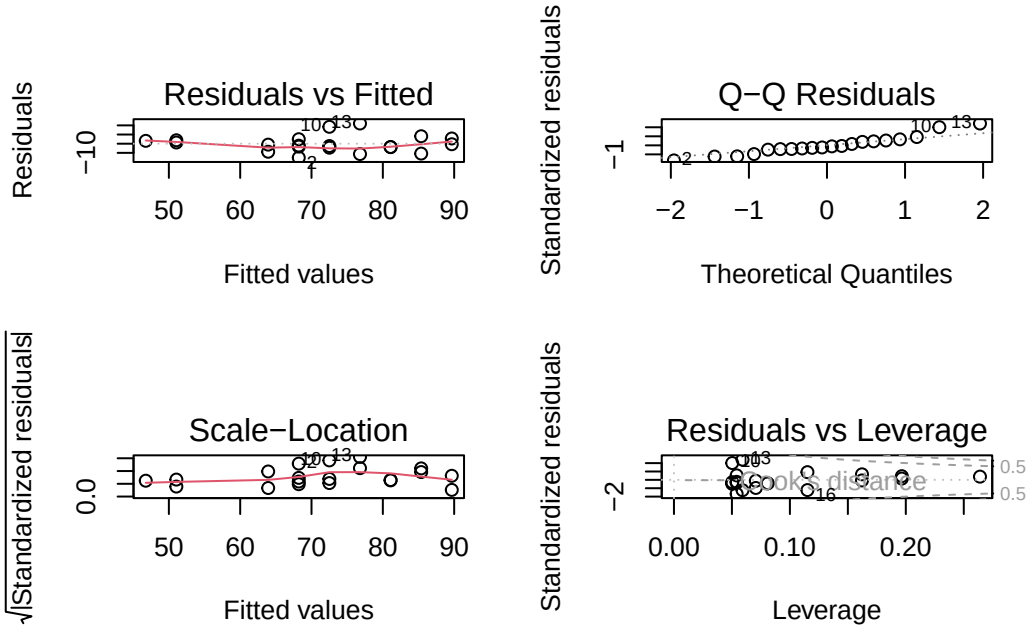


Figure 7.3: Standard regression diagnostic plots for the study-hours and exam-score data (n = 20).

```
par(mfrow = c(1, 1))

coef_display <- summary(model)$coefficients %>%
  as.data.frame() %>%
  tibble::rownames_to_column("Term") %>%
  transmute(
```

```
    Term,
    Estimate = formatC(Estimate, format = "f", digits = 3),
    `Std. error` = formatC(`Std. Error`, format = "f", digits = 3),
    `t value` = formatC(`t value`, format = "f", digits = 2),
    `p value` = formatC(`Pr(>|t|)`, format = "f", digits = 3)
  )

smallsamplelab_apa_table(
  "7.4",
  "Regression coefficients for the simulated diagnostic example",
  coef_display,
  note = sprintf(
    "The fitted model explains %.1f%% of the variance in the outcome
     ↪  (adjusted R^2 = %.3f).",
    100 * summary(model)$adj.r.squared,
    summary(model)$adj.r.squared
  ),
  align = c("l", "r", "r", "r", "r")
)
```

Table 7.4

Regression coefficients for the simulated diagnostic example

Term	Estimate	Std. error	t value	p value
(Intercept)	42.471	5.523	7.69	0.000
study_hours	4.291	0.751	5.71	0.000

Note. The fitted model explains 62.5% of the variance in the outcome (adjusted R^2 = 0.625).

Figure 7.3: Standard regression diagnostic plots for the study-hours and exam-score data (n = 20).

Interpretation: Figure 7.3 brings the standard diagnostics together in one place: the residuals-versus-fitted panel checks linearity, the residual Q-Q plot checks normality of errors, the scale-location panel checks homoscedasticity, and the leverage panel highlights influential cases. Table 7.4 gives the corresponding coefficient estimates, but with small samples a single influential observation can still dominate the fitted line, so sensitivity analysis remains important.

Multicollinearity, Leverage, and Heteroscedasticity

Regression diagnostics in small samples should distinguish three related problems. Multicollinearity means predictors carry overlapping information, which inflates standard errors and makes

individual coefficients unstable. Leverage describes how unusual an observation is in the predictor space. Influence describes how much the fitted coefficients would change if that observation were removed. A case can have high leverage but little influence if it lies close to the fitted regression surface; Cook's distance combines leverage and residual size to flag cases that deserve closer inspection.

Variance inflation factors (VIFs) are a practical first screen for multicollinearity. Values above about 5 warrant inspection, and values above about 10 usually indicate that separate coefficient interpretation is fragile. The condition number gives a whole-model summary of near-linear dependency among predictors; values above about 30 are often treated as a warning sign. These cutoffs are heuristics, but they are especially important when n is small because there are few observations to stabilise correlated predictors.

Table 7.5

Multicollinearity diagnostics for a small regression example

Predictor	Diagnostic	Interpretation
Study hours	5.98	Inspect because it overlaps strongly with practice score
Practice score	6.21	Inspect because it overlaps strongly with study hours
Fatigue	1.09	No obvious collinearity concern
Model condition number	4.8	Whole-model screen for near-linear dependency

Note. VIF values above about 5 warrant inspection and values above about 10 make separate coefficient interpretation fragile. Condition numbers above about 30 are a common whole-model warning sign.

If the residuals-versus-fitted plot shows clear heteroscedasticity, exclusion is usually not the first response. Recheck whether the model is correctly specified, then consider reporting heteroscedasticity-consistent standard errors as a sensitivity check. The `sandwich` and `lmtest` packages provide this workflow:

```
library(sandwich)
library(lmtest)

lmtest::coeftest(vif_model, vcov. = sandwich::vcovHC(vif_model, type =
 ↪  "HC1"))
```

When reporting these diagnostics, avoid vague wording such as "all assumptions were met." A better statement is specific: "Residual plots showed no clear curvature; vertical spread increased slightly at higher fitted values, so HC1 robust standard errors were also computed. The interpretation of the study-hours coefficient did not change." This tells readers what was checked, what was imperfect, and whether the conclusion depended on the diagnostic judgement.

Outlier Disposition Decision Box

When a case is flagged by Mahalanobis distance, leverage, or Cook's distance, first check whether it is a data error. If it is, correct it from the source record or exclude it with a clear audit trail. If the value is genuine but extreme, retain it and report a sensitivity analysis with and without the case. If the case belongs to a different population from the one defined in the research question, exclude it only after documenting why it falls outside the target population. Do not remove a case solely because it weakens statistical significance.

Identifying Data Entry Errors

Data screening should also look for values outside plausible ranges, inconsistent coding schemes, duplicate records, and implausible combinations of attributes. Examples include an age of 150, a Likert response of 8 on a 1–7 scale, one participant ID entered twice, or a primary-school student apparently reporting 20 years of work experience. When any of these appear, cross-check them against source documents or, when appropriate, re-contact participants before deciding whether a correction or exclusion is justified.

Documenting Data Cleaning

Maintain a data-cleaning script that reads the raw data, flags potential outliers or inconsistencies, applies any corrections or exclusions with explicit justifications, and then produces the cleaned dataset used for analysis. In the write-up, report how many observations were excluded, why they were excluded, and how the summary statistics changed after cleaning so readers can see what decisions mattered. When adapting the code in this chapter, record `set.seed()` values and package versions, for example with `sessionInfo()`, so stochastic diagnostics can be reproduced.

Key Takeaways

Data screening matters more, not less, when samples are small because a single unusual observation can distort a mean, a correlation, or a regression coefficient. In practice, the most defensible approach combines visual diagnostics, simple numerical rules, and context-specific judgment rather than mechanically deleting anything extreme. Good screening therefore ends with transparent documentation: what was flagged, what was changed, what remained in the dataset, and whether the conclusions were sensitive to those decisions.

Self-Assessment Quiz

Question 1

Why are outliers particularly problematic in small-sample research?

a) They are easier to detect visually
b) They can have disproportionate influence on results due to the limited number of observations
c) Small samples always have more outliers than large samples
d) Outliers are impossible to detect with small samples

Question 2

What is the primary advantage of Mahalanobis distance over univariate outlier detection methods?

a) It is faster to compute
b) It measures how far an observation lies from the multivariate centre, accounting for covariances among variables
c) It only works with large samples
d) It automatically removes all outliers

Question 3

According to Tukey's fences method, an observation is flagged as a potential outlier if it falls:

a) Within 1.5 × IQR from the quartiles
b) Beyond ±2 standard deviations from the mean
c) Beyond 1.5 × IQR from the quartiles
d) Exactly at the median

Question 4

Why are visual diagnostics (Q-Q plots, boxplots) preferred over formal normality tests (Shapiro–Wilk) for small samples?

a) Visual diagnostics are more statistically rigorous
b) Normality tests have low power with small samples and may not detect departures; visual checks provide more insight
c) Normality tests are too expensive
d) Visual diagnostics automatically calculate p-values

Question 5

In a regression diagnostic plot of "Residuals vs Fitted Values," what does a funnel-shaped pattern indicate?

a) Perfect homoscedasticity
b) Heteroscedasticity—residual variance changes with fitted values
c) Normality of residuals
d) High collinearity among predictors

Question 6

What does Cook's distance measure in regression diagnostics?

a) The distance between two data points
b) The influence of each observation on the regression coefficients; high values indicate influential observations
c) The correlation between predictors
d) The degree of multicollinearity

Question 7

When should an outlier be removed from analysis?

a) Always, because outliers are bad
b) Never, because all data points are valuable
c) Only if there is clear evidence of data entry error or the observation does not belong to the target population
d) Whenever it makes the results statistically significant

Question 8

What is an example of a data entry error that should be flagged during data screening?

a) A participant with a high test score
b) A Likert response of 8 on a 1–7 scale
c) A missing value in a survey
d) A participant who completed all questions

Question 9

Why is documenting data cleaning decisions in a reproducible script important?

a) It allows others to verify exclusions and understand how the cleaned dataset was produced
b) It makes the analysis run faster
c) It is required by all statistical software
d) It prevents missing data from occurring

Question 10

In a Q-Q plot, if the points deviate substantially from the diagonal line at the tails, what does this suggest?

a) The data are perfectly normally distributed
b) The data may have skewness or heavy tails
c) The sample size is too large
d) There are no outliers present

Answers and Explanations

Question 1

Answer: b)

Explanation: Outliers matter more when there are few observations because each case contributes a larger share of the information. A single extreme value in a sample of 15 can drastically shift means, correlations, and regression slopes.

Question 2

Answer: b)

Explanation: Mahalanobis distance evaluates how unusual a case is once the variables are considered together rather than one at a time. That matters because a participant can look ordinary on each separate variable but still form an unusual multivariate combination.

Question 3

Answer: c)

Explanation: Tukey's fences flag values below Q1 - 1.5×IQR or above Q3 + 1.5×IQR. These are screening flags, not automatic deletion rules.

Question 4

Answer: b)

Explanation: With only a small number of observations, formal normality tests often have too little power to detect meaningful departures from the model. A Q-Q plot or boxplot lets the analyst see skewness, heavy tails, and unusual points directly instead of relying on a single p-value.

Question 5

Answer: b)

Explanation: A funnel shape indicates that residual spread changes across the fitted-value range. That pattern violates the constant-variance assumption.

Question 6

Answer: b)

Explanation: Cook's distance measures how much the fitted regression coefficients would change if a case were removed. High values identify observations that deserve sensitivity checks.

Question 7

Answer: c)

Explanation: Removal is justified only when the case is a data error or falls outside the target population. Otherwise, document the case and report sensitivity analyses when it materially affects results.

Question 8

Answer: b)

Explanation: A response of 8 on a 1-7 scale is outside the instrument's valid range. That value should be checked against the source record before analysis.

Question 9

Answer: a)

Explanation: A reproducible cleaning script lets readers verify which cases were flagged, corrected, excluded, or retained. That transparency is essential when each small-sample decision can change the result.

Question 10

Answer: b)

Explanation: Tail departures from the diagonal suggest skewness or heavier or lighter tails than expected under normality.

Chapter 8: Handling Missing Data in Small Samples

Learning Objectives

By the end of this chapter, you will be able to distinguish MCAR, MAR, and MNAR mechanisms, describe missingness patterns clearly, judge when complete-case analysis or multiple imputation is defensible in a small sample, and report missing-data decisions with the transparency needed for readers to assess their consequences.

The Challenge of Missing Data in Small Samples

Missing data are common in applied research. Participants skip survey questions, drop out of longitudinal studies, or provide incomplete records. With large samples, modern methods (multiple imputation, full information maximum likelihood) can handle substantial missingness without excessive bias. With small samples, however, missing data pose severe problems. Even a few missing observations can substantially reduce effective sample size and statistical power.

Missing data methods rely on large-sample asymptotics and may be unstable or inappropriate when samples are very small ($n < 30$) or missingness is extensive ($> 20\%$). In such cases, prevention (minimise missingness through careful design) and transparency (report missingness patterns and sensitivity analyses) are more important than sophisticated imputation.

Types of Missingness

The standard missingness framework distinguishes between three mechanisms (Rubin 1987; Buuren 2018). **MCAR (Missing Completely At Random)** means missingness is unrelated to any observed or unobserved variable, as when a survey page disappears because of a random software glitch. This condition is conceptually simple but rare in practice. **MAR (Missing At Random)** means missingness depends on observed variables but not on the missing values themselves once those observed variables are taken into account. For example, older participants may be more likely to skip a technology question, but conditional on age the missingness is otherwise random. **MNAR (Missing Not At Random)** is the hardest case because missingness depends on the unobserved values themselves, as when participants with severe depression are especially likely to drop out. That possibility cannot usually be resolved from the observed data alone and therefore requires sensitivity analysis or explicit modelling assumptions.

Describing Missingness Patterns

Before choosing a handling strategy, describe the pattern of missingness in plain terms. Report how many observations are missing on each variable, whether the missing values are concentrated in particular participants or particular variables, and whether incomplete cases differ from complete cases on observed characteristics that might make MAR more plausible than MCAR.

Example Dataset for Diagnostics

To demonstrate the handling strategies in this chapter, we simulate a small dataset with missing values on `satisfaction` and `performance`. Table 8.1 shows the first ten rows so the missing-data pattern is visible before the formal diagnostics begin.

```
library(tidyverse)

set.seed(2025)

# Realistic age distribution and MAR missingness:
# performance is more likely to be missing for older participants
age_raw     <- round(rnorm(25, mean = 38, sd = 12))
age_vals    <- pmax(18L, pmin(65L, age_raw))
perf_full   <- round(40 + 0.6 * age_vals + rnorm(25, 0, 10))
miss_prob   <- plogis((age_vals - 42) / 10)
perf_obs    <- ifelse(runif(25) < miss_prob * 0.55, NA_real_, perf_full)

study_data <- tibble(
  participant  = 1:25,
  age          = age_vals,
  satisfaction = sample(c(3:7, NA), 25, replace = TRUE,
                        prob = c(0.15, 0.2, 0.25, 0.2, 0.15, 0.05)),
  performance  = perf_obs
)

study_data_display <- study_data %>%
  mutate(across(c(satisfaction, performance), ~ if_else(is.na(.x), "NA",
   ↪  as.character(.x)))) %>%
  slice(1:10)

smallsamplelab_apa_table(
  "8.1",
  "Simulated study dataset with missing values (rows 1 to 10)",
  study_data_display,
  note = "The full simulated dataset contains 25 participants; the excerpt
   ↪  is shown to illustrate the missing values before diagnosis.",
  align = c("r", "r", "r", "r")
)
```

Table 8.1

Simulated study dataset with missing values (rows 1 to 10)

participant	age	satisfaction	performance
1	45	3	54
2	38	7	54
3	47	7	NA
4	53	5	68
5	42	3	72
6	36	5	51
7	43	3	84
8	37	5	49
9	34	4	66
10	46	4	61

Note. The full simulated dataset contains 25 participants; the excerpt is shown to illustrate the missing values before diagnosis.

Testing the MCAR Assumption

Little's MCAR test evaluates whether missingness is consistent with the MCAR mechanism. The test compares observed means across missing-data patterns. A large p-value suggests MCAR is plausible, whereas a small p-value indicates that missingness likely depends on observed data (i.e., not MCAR). With small samples (n < 50), Little's MCAR test has low power to detect meaningful departures from MCAR. A non-significant result does not confirm MCAR. Supplement it with visual inspection of missingness patterns, complete versus incomplete case comparisons, and substantive reasoning about why data might be missing (Graham 2009).

```
# Little's MCAR test
mcar_output <- smalln_mcar_table(study_data)

smallsamplelab_apa_table(
  "8.2",
  "Little's MCAR test for the simulated study dataset",
  mcar_output,
  note = "A larger p-value is more consistent with MCAR, but with small
   ↪  samples this test is only one piece of evidence and does not confirm
   ↪  MCAR. If the naniar package is unavailable, the table records the
   ↪  number of observed missingness patterns and the formal test should
   ↪  be run before final reporting.",
  align = c("r", "r", "r", "r")
)
```

Table 8.2

Little's MCAR test for the simulated study dataset

Statistic	df	p value	Missing patterns
13.309	6	0.038	3

Note. A larger p-value is more consistent with MCAR, but with small samples this test is only one piece of evidence and does not confirm MCAR. If the naniar package is unavailable, the table records the number of observed missingness patterns and the formal test should be run before final reporting.

```
smalln_missing_var_plot(study_data)
```

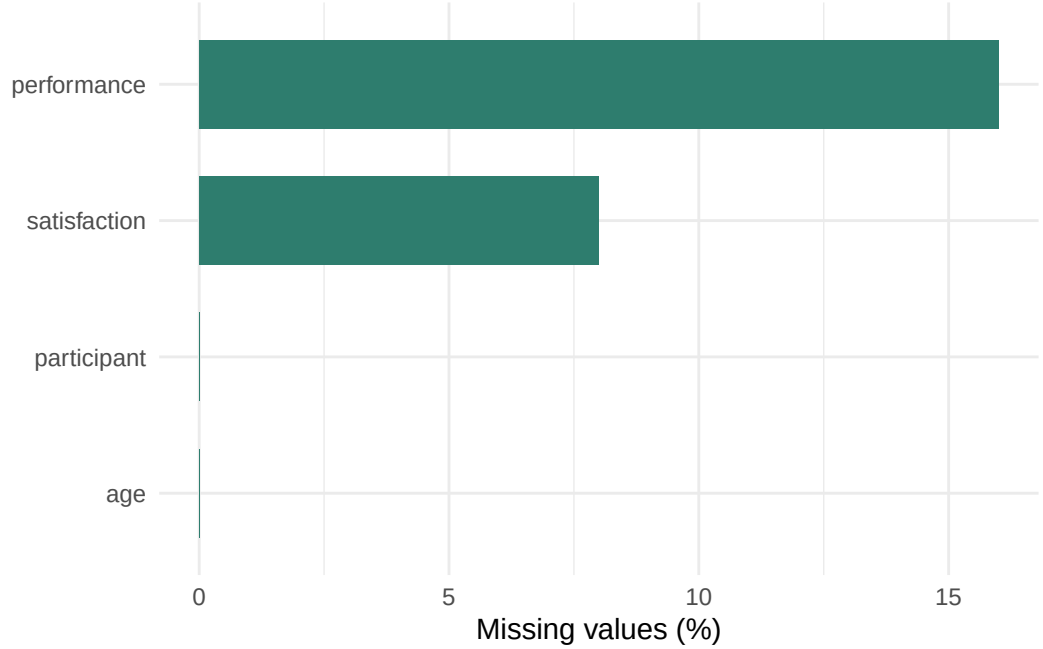


Figure 8.1: Percentage of missing values per variable.

Little's MCAR test assesses whether missing data patterns are completely random. Table 8.2 reports the test statistics, while Figure 8.1 shows the percentage of missing values for each variable. However, with small samples (n < 50), the test has low power and should be treated as one clue rather than a verdict. In practice, it should be supplemented with visual inspection of missingness patterns, comparisons between complete and incomplete cases, and domain knowledge about why participants might have skipped particular items or visits.

Observation-Level Missingness Pattern

Figure 8.2 complements Figure 8.1 by showing whether missing values cluster within particular cases rather than only within particular variables.

```
smalln_missing_matrix_plot(study_data)
```

Figure 8.2: Observation-level missingness pattern.

Example: Summarising Missing Data

We continue working with the simulated dataset (`study_data`) created above. Tables 8.3 to 8.5 summarise how much data are missing and whether the incomplete cases differ from the complete cases on age.

```
# Count missing values per variable
missing_summary <- study_data %>%
  summarise(across(everything(), ~ sum(is.na(.)))) %>%
  pivot_longer(everything(), names_to = "Variable", values_to = "Missing")

smallsamplelab_apa_table(
  "8.3",
  "Number of missing values by variable",
  missing_summary,
```

```
  align = c("l", "r")
)
```

Table 8.3

Number of missing values by variable

Variable	Missing
participant	0
age	0
satisfaction	2
performance	4

```
# Proportion missing
prop_missing <- study_data %>%
  summarise(across(everything(), ~ mean(is.na(.)))) %>%
  pivot_longer(everything(), names_to = "Variable", values_to =
   ↪  "Proportion missing") %>%
  mutate(`Proportion missing` = formatC(`Proportion missing`, format =
   ↪  "f", digits = 2))

smallsamplelab_apa_table(
  "8.4",
  "Proportion of missing values by variable",
  prop_missing,
  align = c("l", "r")
)
```

Table 8.4

Proportion of missing values by variable

Variable	Proportion missing
participant	0.00
age	0.00
satisfaction	0.08
performance	0.16

```
# Compare complete vs incomplete cases
study_data_with_complete <- study_data %>%
  mutate(complete = complete.cases(study_data))

complete_vs_incomplete <- study_data_with_complete %>%
  group_by(complete) %>%
```

```
  summarise(mean_age = mean(age, na.rm = TRUE), n = n(), .groups = "drop")
    %>%
  mutate(
    complete = if_else(complete, "Complete", "Incomplete"),
    mean_age = formatC(mean_age, format = "f", digits = 1)
  ) %>%
  rename(Group = complete, `Mean age` = mean_age, n = n)

smallsamplelab_apa_table(
  "8.5",
  "Age comparison for complete and incomplete cases",
  complete_vs_incomplete,
  align = c("l", "r", "r")
)
```

Table 8.5

Age comparison for complete and incomplete cases

Group	Mean age	n
Incomplete	39.8	6
Complete	41.9	19

Interpretation: Tables 8.3 and 8.4 show that `performance` has the heaviest missingness burden, at about 20% of the sample. Table 8.5 then shows that the incomplete cases are younger on average than the complete cases, which makes MCAR less automatic and suggests that MAR or MNAR should remain plausible possibilities. With 20% missingness and n = 25, only 20 cases remain in a complete-case analysis, so power is reduced sharply.

Complete-Case (Listwise Deletion) Analysis

The simplest approach is to analyse only cases with complete data on all variables of interest. This is valid if missingness is MCAR and the reduction in sample size is tolerable. However, it can introduce bias if missingness is MAR or MNAR, and it wastes information. Complete-case analysis is most defensible when missingness is minimal, MCAR is at least plausible, or the sample is so small that a more elaborate imputation model would be less credible than a transparent analysis of the observed cases.

⚠ Common Misconception: "Listwise Deletion Is Always Safe if Missingness Is Random"

Myth: "If I check for MCAR and the test is non-significant, listwise deletion is unbiased."
Reality: Even when missingness is **truly MCAR**, listwise deletion **loses power** and can introduce bias if you have multiple variables with independent missing patterns.
Demonstration:

```
set.seed(2025)

# Generate complete data: n=50, correlation between x and y = 0.6
n <- 50
x <- rnorm(n, 50, 10)
y <- 0.6 * x + rnorm(n, 0, 8)

# True correlation (no missing data)
true_cor <- cor(x, y)

# Introduce MCAR missingness (20% on x, 20% on y, independently)
x_missing <- x
y_missing <- y
x_missing[sample(1:n, 10)] <- NA  # 20% missing
y_missing[sample(1:n, 10)] <- NA  # 20% missing

# Listwise deletion: only cases with both x and y
complete_cases <- complete.cases(x_missing, y_missing)

# Correlation with listwise deletion
listwise_cor <- cor(x_missing[complete_cases],
 ↪  y_missing[complete_cases])

listwise_demo <- tibble(
  Metric = c(
    "True correlation (complete data)",
    "Complete cases retained",
    "Correlation after listwise deletion",
    "Cases lost",
    "Standard error inflation"
  ),
  Value = c(
    formatC(true_cor, format = "f", digits = 3),
    sprintf("%d / %d (%.1f%%)", sum(complete_cases), n, 100 *
     ↪  mean(complete_cases)),
    formatC(listwise_cor, format = "f", digits = 3),
    sprintf("%d (%.1f%%)", n - sum(complete_cases), 100 * (1 -
     ↪  mean(complete_cases))),
    sprintf("%.2f×", sqrt(1 / sum(complete_cases)) / sqrt(1 / n))
```

```
  )
)

smallsamplelab_apa_table(
  "8.6",
  "Consequences of listwise deletion under independent MCAR
   ↪  missingness",
  listwise_demo,
  note = "With 20% missing on x and 20% missing on y, independent MCAR
   ↪  retention is (1 - p1) x (1 - p2) = 0.8 x 0.8 = 0.64, so only
   ↪  about 64% of cases remain once both variables are required.",
  align = c("l", "l")
)
```

Table 8.6

Consequences of listwise deletion under independent MCAR missingness

Metric	Value
True correlation (complete data)	0.549
Complete cases retained	33 / 50 (66.0%)
Correlation after listwise deletion	0.548
Cases lost	17 (34.0%)
Standard error inflation	1.23×

Note. With 20% missing on x and 20% missing on y, independent MCAR retention is (1 - p1) x (1 - p2) = 0.8 x 0.8 = 0.64, so only about 64% of cases remain once both variables are required.
Why this matters:

1. **Power loss**: With 20% missing on x and 20% on y independently, the retention rate is (1 - p_x) x (1 - p_y) = 0.8 x 0.8 = 0.64, so about 36% of cases are lost.
2. **Multiple variables compound**: With five variables each 15% missing, the retention rate is 0.85^5 = 0.44, so less than half the sample remains.
3. **Bias can still occur**: If missingness is MAR rather than MCAR, listwise deletion can bias estimates as well as reduce precision.

Lesson:

- **MCAR does NOT mean listwise deletion is optimal**—it can still waste information.
- Consider **multiple imputation** when missingness exceeds about 10%, the sample is large enough for the imputation model, and MAR is plausible.
- With small samples (n < 50), losing even 20% of cases can sharply reduce power.

When listwise deletion is actually safe:

- Missingness < 5% on any variable
- n is large enough that losing cases doesn't hurt power
- You've verified MCAR (not just MAR) AND documented the power loss

Mean Imputation (Not Recommended)

Mean imputation replaces missing values with the variable mean. This approach artificially reduces variance and distorts correlations. It is generally not recommended, especially with small samples where each imputed value has disproportionate impact. At most, it might be tolerated for trivial descriptive summaries in a much larger dataset, but it should not be treated as an inferential solution.

Last Observation Carried Forward (LOCF)

In longitudinal studies, LOCF replaces missing follow-up values with the last observed value for that individual. This assumes no change after the last observation, which is often unrealistic. LOCF can therefore bias estimates and is not generally recommended. It is only marginally defensible when the assumption of no meaningful change is substantively plausible and no better alternative is available.

Multiple Imputation (Caution with Small Samples)

Multiple imputation (MI) generates several plausible imputed datasets, analyses each separately, and pools results to account for imputation uncertainty. MI is a strong default for handling missing data in adequately sized samples when MAR is plausible (Rubin 1987; White, Royston, and Wood 2011; Buuren 2018). If MNAR is suspected, MI alone does not solve the problem. The practical response is sensitivity analysis that varies the assumed departure from MAR (Sterne et al. 2009). MI requires sufficient data to estimate imputation models reliably. With very small samples (n < 30) or extensive missingness (> 20%), imputation models may be under-identified and can yield unstable or implausible imputations. If MI is attempted in that setting, use predictive mean matching (`method = "pmm"`), limit the number of predictors in the imputation model, check convergence diagnostics carefully, and report complete-case results as a sensitivity check.

Example: Multiple Imputation with mice (Caution)

We apply MI to the dataset with missing `satisfaction` and `performance` values. Given the small sample (n = 25) and 20% missingness, interpret results cautiously. Table 8.7 reports the pooled regression results rather than printing the raw imputation object.

```
# Multiple imputation requires the 'mice' package
if (requireNamespace("mice", quietly = TRUE)) {
  library(mice)

  # Remove 'complete' indicator variable before imputation
  impute_data <- study_data %>% select(participant, age, satisfaction,
↪  performance)

  # Perform multiple imputation (m = 5 imputations)
  set.seed(2025)
  imp <- mice(impute_data, m = 5, method = "pmm", printFlag = FALSE)

  # Convergence diagnostic to run before reporting:
  # plot(imp)

  # Example analysis: regress performance on age and satisfaction
  fit <- with(imp, lm(performance ~ age + satisfaction))
  pooled <- pool(fit)

  pooled_display <- summary(pooled) %>%
    as_tibble() %>%
    transmute(
      Term = term,
      Estimate = formatC(estimate, format = "f", digits = 3),
      `Std. error` = formatC(std.error, format = "f", digits = 3),
```

```
      Statistic = formatC(statistic, format = "f", digits = 2),
      `p value` = formatC(p.value, format = "f", digits = 3)
    )

  smallsamplelab_apa_table(
    "8.7",
    "Pooled regression results from the illustrative multiple-imputation
     ↪  analysis",
    pooled_display,
    note = "Five predictive-mean-matching imputations were pooled using
     ↪  Rubin's rules. With only 25 cases and 20% missingness, these
     ↪  estimates should be treated as provisional.",
    align = c("l", "r", "r", "r", "r")
  )

} else {
  smallsamplelab_apa_table(
    "8.7",
    "Multiple-imputation availability note",
    tibble(Note = "Install the mice package to run multiple imputation.
     ↪  With very small samples, MI may be unstable; consider
     ↪  complete-case sensitivity analysis."),
    align = "l"
  )
}
```

Table 8.7

Pooled regression results from the illustrative multiple-imputation analysis

Term	Estimate	Std. error	Statistic	p value
(Intercept)	55.443	11.716	4.73	0.000
age	0.566	0.186	3.05	0.007
satisfaction	-3.379	1.608	-2.10	0.054

Note. Five predictive-mean-matching imputations were pooled using Rubin's rules. With only 25 cases and 20% missingness, these estimates should be treated as provisional.

Interpretation: MI generates plausible values for missing data based on observed relationships. The pooled results combine estimates across imputations, with standard errors adjusted for imputation uncertainty. However, with n = 25 and 20% missingness, the imputation model is estimated from limited data, and results may be unstable. Before reporting the pooled estimates, inspect the standard `mice` trace plots with `plot(imp)` and increase `maxit` or simplify the imputation model if the chains drift rather than forming a stable fuzzy pattern. Compare MI

results to complete-case analysis. If they differ substantially, report both and acknowledge uncertainty. Record the random seed, imputation method, number of imputations, and package versions whenever stochastic imputation code is adapted.

Checking Convergence of Multiple Imputation

When using `mice`, always check whether the imputation algorithm has converged. Poor convergence means the imputed values may not be stable, especially with small samples or complex missing data patterns. This chapter keeps the emphasis on handling decisions rather than diagnostic graphics, so the detailed trace-plot and strip-plot workflow is taken up in Chapter 9: For the present chapter, the practical takeaway is that any MI analysis should be accompanied by those diagnostics before it is reported as credible.

Sensitivity Analyses

When missingness is substantial or MNAR is suspected, conduct sensitivity analyses rather than presenting a single imputed answer as definitive. Compare complete-case results with imputed results, vary the assumptions about the missing-data mechanism where possible, and report how much the substantive conclusions change across those scenarios.

Preventing Missing Data

The best approach to missing data is still prevention. Clear instruments, low respondent burden, follow-up for missed appointments or skipped questions, pilot testing of confusing procedures, and good rapport with participants all reduce the need for heroic statistical repair later.

Key Takeaways

Missing-data work in small samples begins with description, not imputation. Researchers need to know how much is missing, where it is missing, and which missingness mechanisms are plausible before choosing a handling strategy. Complete-case analysis can be acceptable in narrowly defined situations but often wastes too much information, while multiple imputation is only as credible as the sample size, missingness level, and modelling assumptions allow. That is why prevention, diagnostics, and sensitivity analysis matter as much as the final pooled estimate.

Self-Assessment Quiz

Test your understanding of missing-data decisions in Chapter 8.

Question 1

What is the key distinction between MCAR and MAR?

a) MCAR means no data are missing; MAR means some data are missing
b) MCAR means missingness is unrelated to observed or unobserved data, whereas MAR allows missingness to depend on observed variables
c) MCAR applies only to experiments; MAR applies only to surveys
d) There is no practical difference between them

Question 2

Why can listwise deletion still be a poor choice even when MCAR is plausible?

a) It always creates impossible values
b) It can throw away many cases and sharply reduce power, especially when missingness occurs on multiple variables
c) It is only allowed in Bayesian analysis
d) It automatically changes MAR data into MNAR data

Question 3

Why is mean imputation generally not recommended?

a) It is too computationally expensive
b) It artificially reduces variance and distorts associations between variables
c) It requires a larger sample than multiple imputation
d) It only works for binary outcomes

Question 4

When is multiple imputation most defensible in this chapter's guidance?

a) When n is moderate, missingness is not too extensive, and MAR is plausible
b) Whenever a dataset contains any missing value at all
c) Only when data are MNAR
d) Only when the outcome is binary

Question 5

Why is last observation carried forward (LOCF) usually a weak solution to missing follow-up data?

a) Because it assumes no meaningful change after the last observed value
b) Because it can only be used with binary outcomes
c) Because it always increases statistical power
d) Because it requires Bayesian software

Question 6

What does Little's MCAR test evaluate?

a) Whether missingness is consistent with a completely-random mechanism
b) Whether multiple imputation has converged
c) Whether the outcome is normally distributed
d) Whether LOCF is acceptable

Question 7

Why are sensitivity analyses important for missing-data work?

a) They prove which missingness mechanism is correct
b) They show whether conclusions depend heavily on the assumptions used to handle missing data
c) They eliminate the need to describe missingness patterns
d) They allow you to ignore complete-case results

Question 8

What is the best overall strategy for dealing with missing data in small-sample studies?

a) Prevent as much missingness as possible through study design and follow-up
b) Default to mean imputation because it is simple
c) Always drop incomplete cases regardless of context
d) Assume MCAR unless the sample is very large

Answers and Explanations

Question 1

Answer: b)

Explanation: The chapter defines MCAR as missingness unrelated to any variables and MAR as missingness related to observed variables but not the missing values themselves. This distinction matters because most modern missing-data methods assume MAR, not MCAR.

Question 2

Answer: b)

Explanation: The chapter's misconception box shows that independent missingness across variables compounds quickly. Even under MCAR, listwise deletion can waste a large fraction of a small dataset and make estimates much less precise.

Question 3

Answer: b)

Explanation: Mean imputation replaces uncertainty with a constant value. The chapter warns that this shrinks variability and biases correlations, which is especially damaging when each observation matters.

Question 4

Answer: a)

Explanation: The chapter describes multiple imputation as most appropriate when the sample is not extremely small, missingness is moderate, and the MAR assumption is plausible. It explicitly cautions that MI can be unstable with $n < 30$ or heavy missingness.

Question 5

Answer: a)

Explanation: The chapter warns that LOCF assumes the participant would have stayed unchanged after the last observed value. That assumption is often unrealistic, so LOCF can bias treatment effects or longitudinal trends.

Question 6

Answer: a)

Explanation: Little's MCAR test compares observed means across missing-data patterns to assess whether the data are consistent with MCAR. The chapter also stresses that, with small samples, this test is only one clue rather than a final verdict.

Question 7

Answer: b)

Explanation: Because the missingness mechanism is often uncertain, the chapter recommends comparing results under different reasonable assumptions. If results change materially, that uncertainty should be reported rather than hidden.

Question 8

Answer: a)

Explanation: The chapter ends by stressing prevention: clear instruments, reduced burden, follow-up procedures, and pilot testing. No statistical fix can fully recover information that was never observed.

Chapter 9: Assessing Multiple Imputation Quality

Learning Objectives

By the end of this chapter, you will be able to explain why multiple-imputation diagnostics matter, inspect the main convergence and plausibility checks produced by `mice`, evaluate whether pooled estimates are stable across different values of `m`, and report imputation diagnostics in a way that makes downstream analyses defensible.

Why Imputation Diagnostics Matter

Multiple imputation (MI) requires deliberate specification at each stage. The quality of the imputed values depends on whether the imputation models are specified sensibly, whether the chained equations have converged, whether the imputed values remain plausible relative to the observed data, and whether the pooled estimates are stable across different choices of `m`. If those checks are ignored, the result can be biased parameter estimates, incorrect standard errors, implausible imputations, and misplaced confidence in apparently polished output.

Diagnostic 1: Convergence Checks

The `mice` algorithm uses **iterative chained equations**: it cycles through variables, updating imputations based on the current values of other variables. Convergence occurs when these iterations stabilise (no systematic trends).

Trace Plots

Trace plots show the mean and SD of imputed values across iterations for each variable. Good convergence looks like a fuzzy caterpillar: trace lines fluctuate randomly around a stable mean, show no systematic upward or downward trend across iterations, and chains from different imputations intermingle rather than remaining separated. Figure 9.1 shows the full set of trace plots for the three simulated variables.

```r
library(mice)
library(tidyverse)

# Simulate data with missing values
set.seed(2025)
mi_data <- tibble(
  age = c(25, 32, NA, 45, 29, NA, 38, 41, 27, 35, NA, 42, 30, 28, 39),
  satisfaction = c(4, 5, 3, NA, 4, 5, NA, 4, 5, 3, 4, NA, 5, 4, 3),
  income = c(35, 50, 42, 60, NA, 55, 48, NA, 40, 52, 45, 58, NA, 38, 49)
)

# Multiple imputation with more iterations to demonstrate convergence
imp <- mice(mi_data, m = 5, maxit = 20, seed = 2025, print = FALSE)

# Plot trace lines for all variables
plot(imp, c("age", "satisfaction", "income"))
```

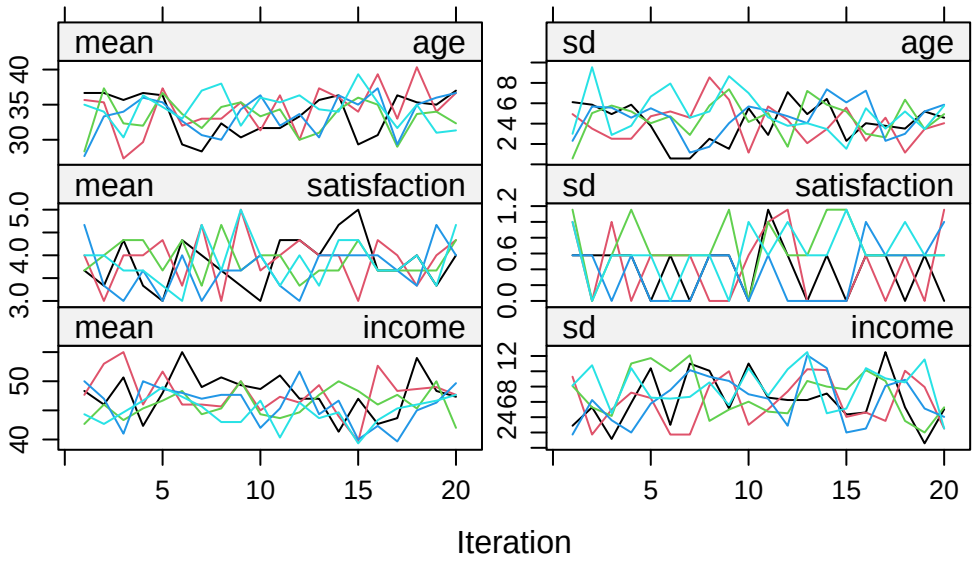


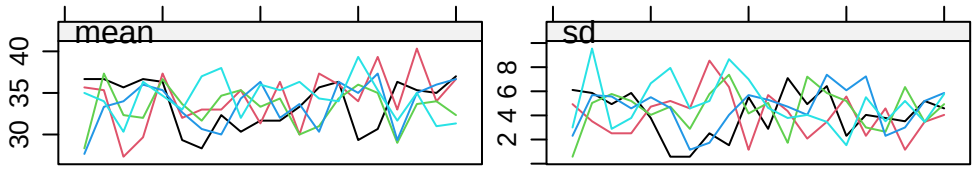
Figure 9.1: Trace plots for age, satisfaction, and income across imputations.

Interpretation: Figure 9.1 should resemble a fuzzy caterpillar rather than a drifting set of lines. If the traces still drift after roughly 20 iterations, increase `maxit`. If chains remain separated, inspect the imputation model specification rather than assuming convergence has occurred.

Checking Specific Variables

If you have many variables, focus on those with the most missingness. Figure 9.2 shows how that targeted check looks when attention is restricted to `age`.

```
# Focus on specific variables with high missingness
plot(imp, "age")
```



Iteration

Figure 9.2: Trace plot for age only.

If you still see clear trends after the first 10 to 20 iterations, increase `maxit` to 50 or even 100. In many routine MCAR or MAR settings, `maxit = 20-50` is usually enough, but the diagnostics should drive that decision rather than a hard default.

Diagnostic 2: Imputed vs. Observed Distributions

Imputed values should **resemble** the observed data distribution (but not be identical). Large discrepancies suggest model misspecification. With n < 30, imputed distributions may appear jagged or narrower because predictive mean matching has few donor values. Focus on whether imputed values fall within a plausible observed range rather than demanding smooth density curves. Figure 9.3 shows the density comparison for the simulated example.

Density Plots

```
# Compare density plots: blue = observed, red = imputed
densityplot(imp)
```

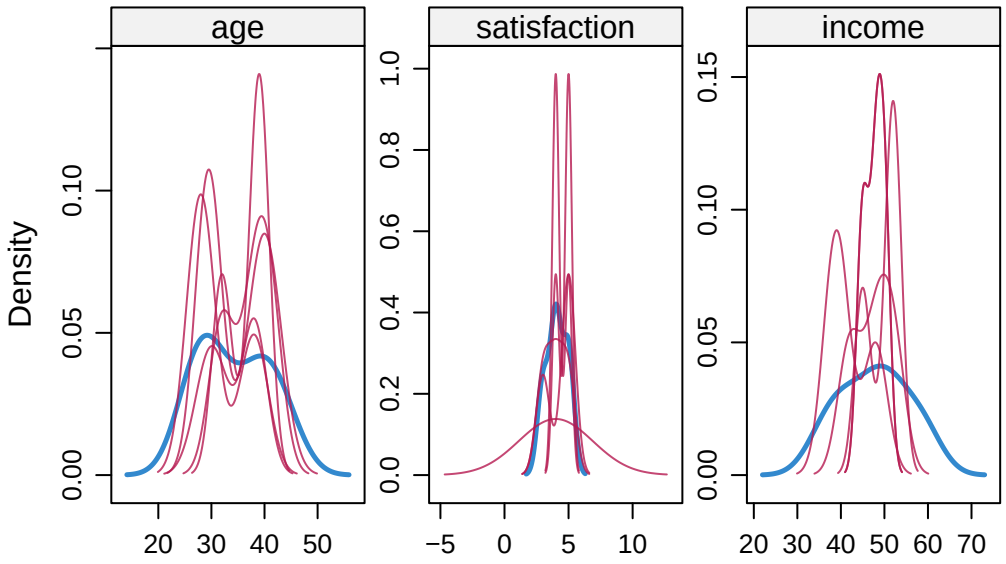


Figure 9.3: Observed and imputed density plots for the simulated example.

Interpretation: Figure 9.3 should show substantial overlap between observed and imputed distributions without making them identical. In very small examples the red curve can look more jagged or narrower because it is based on few imputed values. The red flags are collapse toward a single implausible value or a clear shift outside the observed range.

Strip Plots (Univariate)

Strip plots show individual imputed values (red) alongside observed values (blue). Figure 9.4 uses `age` as the example variable.

```
# Strip plots for each variable
stripplot(imp, age ~ .imp, pch = 20, cex = 1.5)
```

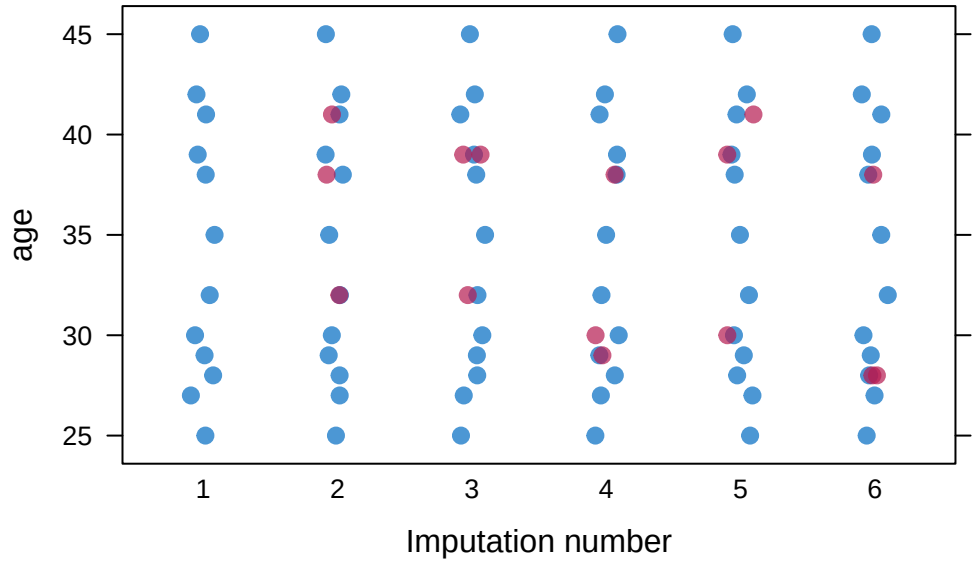


Figure 9.4: Strip plot for age across imputations.

Interpretation: Figure 9.4 should show the imputed values filling gaps in the observed data without behaving like a separate distribution. A useful question here is whether any imputed points fall implausibly far outside the observed range, because that usually signals a poor imputation model rather than legitimate uncertainty.

Diagnostic 3: Sensitivity to m (Number of Imputations)

The number of imputations (m) affects the precision of pooled estimates. With more imputations, pooled estimates become more stable and standard errors more accurate. Table 9.1 compares the coefficient and standard-error estimates across three values of `m`.

Rule of Thumb for m

The number of imputations should increase as the fraction of missing information (FMI) increases. When FMI is below about 10%, `m = 5` to `10` is often adequate. When FMI is around 10% to 30%, `m = 20` to `50` is more defensible. When FMI exceeds 30%, values such as `m = 50` to `100` may be needed. A useful heuristic from White, Royston, and Wood (2011) (White, Royston, and Wood 2011) is $m \geq 100 \times \text{FMI}$, where FMI is averaged across the parameters that matter for the analysis. For example, average FMI = 0.15 suggests at least 15 imputations, and average FMI = 0.30 suggests at least 30. Round up to convenient reporting values such as `m = 20` or `m = 50`.

Testing Sensitivity

```r
# Simulate larger dataset for demonstration
set.seed(2025)
age_base     <- round(rnorm(50, mean = 38, sd = 10))
age_vals50   <- pmax(20L, pmin(60L, age_base))
income_vals  <- round(25 + 0.5 * age_vals50 + rnorm(50, 0, 8))
sat_cont     <- 1 + 0.04 * income_vals + rnorm(50, 0, 0.8)
sat_vals     <- pmin(5L, pmax(1L, round(sat_cont)))

# MAR: satisfaction missing for higher-age participants
miss_prob50  <- plogis((age_vals50 - 42) / 8)
sat_obs      <- ifelse(runif(50) < miss_prob50 * 0.4, NA_real_, sat_vals)

# Small MCAR fraction on age and income for realistic imputation demo
age_obs      <- ifelse(runif(50) < 0.10, NA_real_, age_vals50)
income_obs   <- ifelse(runif(50) < 0.06, NA_real_, income_vals)

mi_data_large <- tibble(
  age          = age_obs,
  income       = income_obs,
  satisfaction = sat_obs
)

# Impute with varying m
imp_m5 <- mice(mi_data_large, m = 5, maxit = 20, seed = 2025, print =
 ↪  FALSE)
imp_m20 <- mice(mi_data_large, m = 20, maxit = 20, seed = 2025, print =
 ↪  FALSE)
imp_m50 <- mice(mi_data_large, m = 50, maxit = 20, seed = 2025, print =
 ↪  FALSE)

# Fit model and pool results
fit_m5 <- with(imp_m5, lm(satisfaction ~ age + income))
fit_m20 <- with(imp_m20, lm(satisfaction ~ age + income))
fit_m50 <- with(imp_m50, lm(satisfaction ~ age + income))

pooled_m5 <- pool(fit_m5)
pooled_m20 <- pool(fit_m20)
pooled_m50 <- pool(fit_m50)

# Compare coefficient estimates and SEs
compare_m <- tibble(
  m = c(5, 20, 50),
  age_coef = c(
    summary(pooled_m5)$estimate[2],
    summary(pooled_m20)$estimate[2],
    summary(pooled_m50)$estimate[2]
```

```
  ),
  age_se = c(
    summary(pooled_m5)$std.error[2],
    summary(pooled_m20)$std.error[2],
    summary(pooled_m50)$std.error[2]
  )
)

compare_m_display <- compare_m %>%
  transmute(
    m,
    `Age coefficient` = formatC(age_coef, format = "f", digits = 3),
    `Age SE` = formatC(age_se, format = "f", digits = 3)
  )

smallsamplelab_apa_table(
  "9.1",
  "Sensitivity of the age coefficient to the number of imputations",
  compare_m_display,
  note = "Only modest changes across m values are expected once Monte
   ↪  Carlo error is under control.",
  align = c("r", "r", "r")
)
```

Table 9.1

Sensitivity of the age coefficient to the number of imputations

m	Age coefficient	Age SE
5	-0.022	0.013
20	-0.021	0.015
50	-0.019	0.014

Note. Only modest changes across m values are expected once Monte Carlo error is under control.

Interpretation: Table 9.1 should show broadly similar coefficient estimates across different values of `m`, with only modest differences due to Monte Carlo error, and the standard errors should stabilise as `m` increases. If the coefficients move substantially, for example by more than about 10%, that is a sign to increase `m` rather than treating the smaller-imputation result as settled.

When to Use Larger m

Larger values of `m` are especially sensible when missingness is high, when the sample itself is small and Monte Carlo error is therefore more noticeable, or when the analysis is sensitive enough

that conservative inference matters. In practice, that means using at least `m = 20` once missingness moves beyond about 20%, and considering `m = 50` or more for particularly consequential analyses.

Diagnostic 4: Checking Imputation Model Assumptions

Inspect Imputation Methods

Table 9.2 records which imputation method is being used for each variable.

```
# Check which imputation methods were used
method_display <- tibble(
  Variable = names(imp$method),
  Method = unname(imp$method)
)

smallsamplelab_apa_table(
  "9.2",
  "Imputation methods used for each variable",
  method_display,
  align = c("l", "l")
)
```

Table 9.2

Imputation methods used for each variable

Variable	Method
age	pmm
satisfaction	pmm
income	pmm

Common methods: `pmm` uses predictive mean matching and is usually the safest default for continuous variables because it preserves plausible observed values. `norm` uses Bayesian linear regression and therefore leans harder on normality assumptions. `logreg` is intended for binary variables, while `polyreg` is used for unordered categorical variables. In most small-sample settings, `pmm` is the best starting point for continuous variables unless you have a strong reason to prefer a parametric normal model.

Check Predictor Matrix

Table 9.3 shows the predictor matrix that tells `mice` which variables are used to impute each target variable.

```
# See which variables predict each other
predictor_display <- imp$predictorMatrix %>%
  as.data.frame() %>%
  tibble::rownames_to_column("Imputed variable")

smallsamplelab_apa_table(
  "9.3",
  "Predictor matrix for the illustrative imputation model",
  predictor_display,
  note = "A value of 1 means the column variable is used to predict the
   ↪  row variable.",
  align = c("l", "r", "r", "r")
)
```

Table 9.3

Predictor matrix for the illustrative imputation model

Imputed variable	age	satisfaction	income
age	0	1	1
satisfaction	1	0	1
income	1	1	0

Note. A value of 1 means the column variable is used to predict the row variable.

Interpretation: In the predictor matrix, rows identify the variables being imputed and columns identify the variables used to predict them. A 1 means the column variable is included as a predictor for the row variable, while a 0 means it is excluded.

Modify if needed:

```
# Example: Exclude a variable from predicting another
pred <- imp$predictorMatrix
pred["age", "satisfaction"] <- 0  # Don't use satisfaction to predict age

# Re-run imputation with modified predictor matrix
imp_modified <- mice(mi_data, m = 5, predictorMatrix = pred, print = FALSE)
```

Diagnostic 5: Fraction of Missing Information (FMI)

The FMI quantifies how much uncertainty is introduced by imputation. It is automatically reported by `pool()`, and Table 9.4 shows the relevant columns for the pooled regression model.

```
# Pool results and examine FMI
pooled_result <- pool(fit_m20)
pooled_display <- pooled_result$pooled %>%
  as_tibble() %>%
  transmute(
    Term = term,
    Estimate = formatC(estimate, format = "f", digits = 3),
    `Std. error` = formatC(sqrt(t), format = "f", digits = 3),
    Lambda = formatC(lambda, format = "f", digits = 3),
    FMI = formatC(fmi, format = "f", digits = 3)
  )

smallsamplelab_apa_table(
  "9.4",
  "Pooled regression estimates with lambda and FMI",
  pooled_display,
  align = c("l", "r", "r", "r", "r")
)
```

Table 9.4

Pooled regression estimates with lambda and FMI

Term	Estimate	Std. error	Lambda	FMI
(Intercept)	0.101	0.643	0.296	0.342
age	-0.021	0.015	0.214	0.258
income	0.075	0.015	0.232	0.276

Columns to examine: The `fmi` column reports the fraction of missing information for each coefficient, while `lambda` shows the proportion of total variance attributable to missingness.

Interpretation: Values of `fmi` below about `0.10` indicate low missing-information burden and are often compatible with `m = 5` to `10`. Values around `0.10` to `0.30` suggest a moderate burden and support using `m = 20` to `50`. Once `fmi` exceeds about `0.30`, missingness is making a large contribution to uncertainty. Consider larger `m`, add defensible auxiliary variables, or state that MI may not fully recover information in a very small sample.

Example: Full Diagnostic Workflow

The earlier sections show the individual diagnostics. This closing example condenses them into a reporting workflow so the end result is not a second copy of the same plots, but a compact summary of what should be written up after those checks have been completed.

```
library(naniar)

missing_summary_final <- miss_var_summary(mi_data_large) %>%
  transmute(
    Variable = variable,
    `Missing n` = n_miss,
    `Missing %` = formatC(pct_miss, format = "f", digits = 1)
  )

imp_final <- mice(mi_data_large, m = 20, maxit = 30, seed = 2025, print =
 ↪  FALSE)
fit_final <- with(imp_final, lm(satisfaction ~ age + income))
pooled_final <- pool(fit_final)

pooled_final_summary <- pooled_final$pooled %>%
  as_tibble()

report_workflow <- tibble(
  Step = c(
    "Describe missingness",
    "Choose imputation settings",
    "Check convergence",
    "Check plausibility",
    "Pool model estimates",
    "Summarise FMI",
    "Write the report"
  ),
  Action = c(
    paste(missing_summary_final$Variable, missing_summary_final$`Missing
     ↪  %`, "% missing", collapse = "; "),
    "Use predictive mean matching with m = 20, maxit = 30, and random seed
     ↪  = 2025.",
    "Inspect trace plots for drift or separated chains.",
    "Compare observed and imputed distributions with density and strip
     ↪  plots.",
    "Pool the regression of satisfaction on age and income.",
    sprintf(
      "FMI ranges from %s to %s.",
      formatC(min(pooled_final_summary$fmi), format = "f", digits = 3),
      formatC(max(pooled_final_summary$fmi), format = "f", digits = 3)
    ),
    sprintf(
```

```
      "Example write-up: We used predictive mean matching with m = 20
       ↪  imputations, maxit = 30, and random seed = 2025. Diagnostic
       ↪  plots indicated adequate convergence and plausible imputations,
       ↪  and FMI values ranged from %s to %s.",
      formatC(min(pooled_final_summary$fmi), format = "f", digits = 3),
      formatC(max(pooled_final_summary$fmi), format = "f", digits = 3)
    )
  )
)

smallsamplelab_apa_table(
  "9.5",
  "Workflow summary for assessing multiple-imputation quality",
  report_workflow,
  align = c("l", "l")
)
```

Table 9.5

Workflow summary for assessing multiple-imputation quality

Step	Action
Describe missingness	satisfaction 20.0 % missing; age 10.0 % missing; income 6.0 % missing
Choose imputation settings	Use predictive mean matching with m = 20, maxit = 30, and random seed = 2025.
Check convergence	Inspect trace plots for drift or separated chains.
Check plausibility	Compare observed and imputed distributions with density and strip plots.
Pool model estimates	Pool the regression of satisfaction on age and income.
Summarise FMI	FMI ranges from 0.208 to 0.302.
Write the report	Example write-up: We used predictive mean matching with m = 20 imputations, maxit = 30, and random seed = 2025. Diagnostic plots indicated adequate convergence and plausible imputations, and FMI values ranged from 0.208 to 0.302.

Red Flags and Troubleshooting

Problem	Symptom	Solution
Non-convergence	Trace plots show trends	Increase `maxit` (try 50–100)
Imputed values at one value	Density plot shows spike	Use `method = "pmm"` instead of `norm`
Imputed values out of range	Strip plot shows outliers	Check variable type (e.g., use `logreg` for binary)
Unstable estimates across m	Coefficients vary > 10%	Increase m (try 50–100)
High FMI (> 0.50)	Large uncertainty	Consider whether MI is appropriate; may need auxiliary variables or accept wider CIs
Separation warnings (logistic regression)	Model fails to converge	Use penalized imputation methods or increase sample size

Reporting MI Diagnostics

When reporting MI results, include:

1. **Missingness pattern**: "Three variables had missing data (age: 20%, income: 18%, satisfaction: 10%)"
2. **Imputation model**: "We used predictive mean matching with m = 20 imputations, maxit = 30, and random seed = 2025"
3. **Convergence**: "Trace plots showed convergence after 20 iterations (see Supplementary Figure S1)"
4. **Plausibility**: "Imputed values were visually consistent with observed distributions (density plots in Supplementary Figure S2)"
5. **Sensitivity**: "Results were stable across m = 5, 20, and 50 imputations (coefficient differences < 5%)"
6. **FMI**: "Fraction of missing information ranged from 0.12 to 0.25, indicating moderate impact of missingness"

Key Takeaways

Multiple-imputation results are only as defensible as the diagnostics behind them. Convergence checks, distributional comparisons, stability across different values of `m`, and inspection of FMI

all help determine whether the imputation model is behaving plausibly rather than merely producing polished output. In small samples especially, report the seed, `m`, `maxit`, method, package versions, and remaining uncertainty clearly.

Self-Assessment Quiz

Question 1

What is the main purpose of a trace plot in multiple-imputation diagnostics?

a) To check whether the imputation chains have stabilised across iterations
b) To count the number of missing values in each variable
c) To show the final pooled regression coefficients
d) To choose between complete-case analysis and LOCF

Question 2

What pattern in a trace plot would most strongly suggest that `maxit` should be increased?

a) A systematic upward or downward drift across iterations
b) Random fluctuation around a stable mean
c) Several chains overlapping each other
d) Slightly different starting values for each chain

Question 3

What is the main diagnostic question answered by a density plot of observed and imputed values?

a) Whether the imputed values are plausible relative to the observed distribution
b) Whether the predictor matrix contains enough zeros
c) Whether the sample size exceeds 100
d) Whether Little's MCAR test is significant

Question 4

In a strip plot, what would count as a red flag?

a) Imputed values falling well outside the observed range
b) Observed and imputed points using different colors
c) The x-axis being labelled by imputation number
d) Having more than one imputed dataset

Question 5

Why does the chapter compare pooled results across `m = 5`, `20`, and `50`?

a) To check whether the estimates are stable as Monte Carlo error is reduced
b) To prove that larger m always changes the coefficient direction
c) To replace the need for FMI
d) To test whether the outcome is normally distributed

Question 6

What is the rule of thumb from White, Royston, and Wood (2011) for choosing `m`?

a) Use m ≥ 100 × FMI
b) Always use m = 5
c) Set m equal to the sample size
d) Use m = 2 whenever missingness is below 10%

Question 7

What does the predictor matrix tell you in a `mice` analysis?

a) Which variables are used to predict each imputed variable
b) Which cases were dropped before imputation
c) How many imputations are needed
d) Whether the pooled p-values are significant

Question 8

Why is predictive mean matching (`pmm`) often preferred to `norm` for continuous variables in small samples?

a) Because `pmm` preserves plausible observed values and is less dependent on strict normality
b) Because `pmm` requires no predictors
c) Because `norm` only works for binary data
d) Because `pmm` guarantees narrower confidence intervals

Question 9

What does a high FMI value mean for a pooled coefficient?

a) That missingness is contributing substantially to uncertainty in that estimate
b) That the predictor matrix is incorrect
c) That the data are definitely MNAR
d) That convergence has failed

Question 10

What should you do if the MI results differ materially from the complete-case results?

a) Report both and discuss the sensitivity of the conclusion to the missing-data assumptions
b) Automatically prefer the imputed result
c) Delete the complete-case analysis from the paper
d) Increase the significance threshold

Question 11

What information should appear in a transparent write-up of an MI analysis?

a) The imputation method, the number of imputations, the convergence/plausibility checks, and the FMI summary
b) Only the final p-values from the pooled model
c) Only the percentage of missing values
d) Only the complete-case results

Answers and Explanations

Question 1

Answer: a)

Explanation: Trace plots are a convergence diagnostic. They show whether the imputed values are fluctuating around a stable level or still drifting across iterations.

Question 2

Answer: a)

Explanation: If the trace continues to trend upward or downward, the chained-equations algorithm has not yet stabilised. That is the clearest signal to increase `maxit`.

Question 3

Answer: a)

Explanation: Density plots compare the observed and imputed distributions. The goal is to check plausibility, not to force the two distributions to be identical.

Question 4

Answer: a)

Explanation: Strip plots should show imputed values filling plausible gaps in the observed data. Values far outside the observed range usually indicate model misspecification.

Question 5

Answer: a)

Explanation: Changing `m` is a sensitivity check for Monte Carlo error. If the pooled estimates move substantially, then the smaller value of `m` was not yet stable enough.

Question 6

Answer: a)

Explanation: The chapter recommends the shorthand m >= 100 x FMI, which scales the number of imputations to the amount of missing information in the model. For example, an average FMI of 0.30 suggests at least 30 imputations.

Question 7

Answer: a)

Explanation: Rows identify the variable being imputed and columns identify the available predictors. A 1 means the predictor is used for that row variable, and a 0 means it is excluded.

Question 8

Answer: a)

Explanation: The chapter treats `pmm` as the safer default because it draws on observed donor values. That makes it more robust when normal-theory assumptions are not especially convincing.

Question 9

Answer: a)

Explanation: FMI quantifies how much uncertainty is coming from the missing data rather than only from observed-data sampling variation. High FMI values justify larger m and more cautious interpretation.

Question 10

Answer: a)

Explanation: Material disagreement between MI and complete-case results is itself important information. The chapter recommends reporting that sensitivity rather than hiding it.

Question 11

Answer: a)

Explanation: The final reporting section stresses that readers need to know the method, m, the main diagnostics, and the FMI burden. Reporting only the pooled coefficients would hide the quality checks that make the analysis credible.

Summary of Part B

Part B addressed the design, measurement, and data-quality decisions that make small-sample analyses credible. The chapters covered sampling strategies that maximise information with limited resources, measurement quality and scale development, reliability for short scales, short-scale development, data screening and diagnostic checks, and the identification and handling of missing data, including multiple imputation and the assessment of imputation quality.

The main point is that small-sample analysis begins well before the final model is fitted. Transparent sampling, careful measurement, early diagnostic checking, and defensible handling of missing data all determine whether the later inferential results can be trusted.

Part C: Analysis Methods

This part presents the core toolkit for small-sample quantitative analysis. We move from exact and resampling tests to nonparametric rank-based methods, methods for sparse counts and short time series, penalised and Bayesian regression, and multi-criteria decision-making (MCDM).

The companion lab volume contains the guided practicals and challenge activities associated with these chapters. This textbook edition focuses on the conceptual explanation, worked examples, and interpretation needed to choose and report the methods appropriately.

In This Part

- Chapter 10: Exact Tests and Resampling Methods
- Chapter 11: Nonparametric Rank-Based Methods
- Chapter 12: Methods for Sparse Counts and Short Time Series
- Chapter 13: Penalised and Bayesian Regression for Small Samples
- Chapter 14: Multi-Criteria Decision Making (MCDM) for Small Sets of Alternatives
- Summary of Part C

Chapter 10: Exact Tests and Resampling Methods

Small-sample analysis often requires reference distributions that do not depend on large-sample theory.

Exact tests and resampling methods become especially useful when sample sizes are modest and large-sample approximations carry meaningful risk of inaccuracy. This chapter explains when to use exact tests for discrete outcomes, when permutation tests provide a cleaner reference distribution than a parametric model, and when bootstrap resampling is useful for interval estimation. The emphasis is practical: matching the method to the design, the outcome type, and the inferential goal.

Learning Objectives

By the end of this chapter, you will be able to explain when exact tests are preferable to large-sample approximations, distinguish conditional exact tests from unconditional and mid-p alternatives, implement exact binomial, exact Poisson, permutation, and bootstrap procedures in R, and report resampling analyses with enough detail for readers to reproduce the statistic, number of resamples, random seed, and inferential target.

When to Use Exact and Resampling Methods

Exact tests calculate p-values directly from the null distribution of the statistic. They are especially useful when the outcome is discrete, when expected cell counts are small, or when the sample is too limited for asymptotic results to be reliable.

Resampling methods use the observed data to approximate a sampling distribution. Permutation tests reassign labels under the null hypothesis to create a reference distribution for a test statistic. Bootstrap methods resample with replacement from the observed data to approximate the variability of an estimator and to construct confidence intervals.

In practice, exact tests are most natural for sparse binary or count data. Permutation tests are useful when exchangeability under the null is plausible. Bootstrap methods are especially helpful when the statistic of interest lacks a simple closed-form standard error.

For any resampling analysis, report the statistic being resampled and the number of permutations or bootstrap resamples. Also report the random seed and whether the result is exact or Monte Carlo. Those details determine the stability and reproducibility of the reported inference.

Fisher's Exact Test for 2×2 Tables

Fisher's exact test is the standard conditional test for association in a 2×2 contingency table when some expected cell counts are small. It conditions on the observed margins and calculates the probability of the observed table, and all equally or more extreme tables, under the null hypothesis of no association.

Because the sample space is discrete, Fisher's test can be conservative: its p-value is guaranteed to control the Type I error rate, but that guarantee can come at the cost of reduced power. That trade-off is often acceptable in confirmatory work, especially when one of the margins is fixed by design.

Example: Fisher's Exact Test

Suppose we are evaluating a new training intervention. Of 10 employees who received training, 8 met their performance target; of 10 who did not receive training, 3 met the target. We test whether training is associated with meeting the target.

```
training_table <- matrix(
  c(8, 2, 3, 7),
  nrow = 2,
  byrow = TRUE,
  dimnames = list(
    Training = c("Yes", "No"),
    Target = c("Met", "Not Met")
  )
)

training_display_table <- tibble(
  `Training group` = c("Training", "No training"),
  `Met target` = c(8, 3),
  `Did not meet target` = c(2, 7)
)

training_expected_counts <-
 ↪  suppressWarnings(chisq.test(training_table)$expected)
fisher_result <- fisher.test(training_table)

smallsamplelab_apa_table(
  "10.1",
  "Training outcomes by group",
```

```
  training_display_table,
  note = sprintf(
    "The smallest expected cell count under independence is %.1f, so exact
     ↪  inference is preferable to the chi-square approximation.",
    min(training_expected_counts)
  ),
  align = c("l", "r", "r")
)
```

Table 10.1

Training outcomes by group

Training group	Met target	Did not meet target
Training	8	2
No training	3	7

Note. The smallest expected cell count under independence is 4.5, so exact inference is preferable to the chi-square approximation.

Fisher's exact test gives an odds ratio of 8.15, with a 95% confidence interval from 0.88 to 127.06, and an exact two-sided p-value of 0.070. That odds ratio, despite the wide interval, points toward a meaningful association that a larger study would be well placed to assess.

Visualising the 2×2 Table

A mosaic-style plot helps readers see the same 2×2 structure graphically. The two training groups have equal width because each contains 10 employees, and the vertical split within each block shows the proportion who met the target. In this example, the training group has a visibly larger share of employees meeting the target.

```
training_mosaic_plot
```

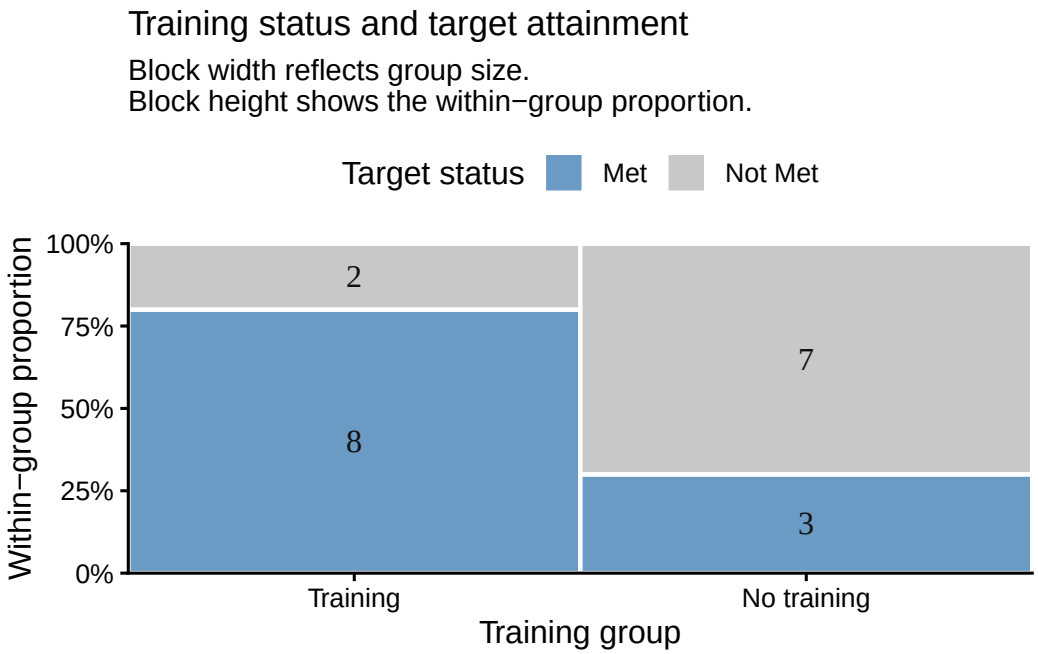


Figure 10.1: Mosaic-style plot of training status and target attainment. Block width reflects group size; block height within each group reflects the proportion meeting the target.

When Fisher's Exact Test Is Conservative

Fisher's exact test conditions on both row and column margins of the 2x2 table. In prospective studies, only the group sizes are fixed by design; conditioning on the observed outcome margin as well can make the test conservative. In case-control studies, both margins may align more closely with the sampling plan, though Fisher's conditioning remains stricter than the design itself.

- **Prospective study or trial**: group sizes are fixed by design, so one margin matches the design directly. Fisher's conditioning still also fixes the observed outcome margin, which is random and can make the test conservative.
- **Case-control study**: case and control totals are fixed by design, so the conditioning aligns more closely with the sampling plan, though Fisher's exact conditioning is still stricter than the design itself.
- **Cross-sectional or convenience sample**: neither margin is fixed by design, so the extra conditioning is most visible and conservatism is often easiest to see.

The practical implication is straightforward: a Fisher p-value can stay above 0.05 even when a less conservative exact method gives stronger evidence. Fisher's test remains valid, but the conditioning scheme affects how much power is available.

Unconditional Exact Tests

The unconditional approach, developed by Barnard, does not require conditioning on both margins. Modern implementations therefore offer **unconditional exact tests** that condition only on the group sizes. These tests are often more powerful than Fisher's exact test when the margins are not fixed by design.

In R, the `exact2x2` package provides unconditional exact procedures through `uncondExact2x2()` (Fay 2010). Install it with `install.packages("exact2x2")` if it is not already available. The score version shown below is Barnard-style in spirit and compares the observed proportions while conditioning only on the group sizes.

```
library(exact2x2)

midp_result <- exact2x2(training_table, alternative = "two.sided", midp =
 ↪  TRUE)
barnard_result <- uncondExact2x2(
  x1 = 8,
  n1 = 10,
  x2 = 3,
  n2 = 10,
  parmtype = "difference",
  alternative = "two.sided",
  method = "score",
  conf.int = TRUE
)
# uncondExact2x2() returns p2 - p1, so reverse the sign to report training
 ↪  minus no training.
barnard_risk_difference <- unname(-barnard_result$estimate)
barnard_risk_difference_ci <- c(-barnard_result$conf.int[2],
 ↪  -barnard_result$conf.int[1])

exact_comparison_table <- tibble(
  Method = c(
    "Fisher's exact test",
    "Fisher's exact test with mid-p correction",
    "Unconditional exact score test"
  ),
  Conditioning = c(
    "Conditions on both margins",
    "Conditions on both margins, then applies a mid-p adjustment",
    "Conditions on group sizes only"
  ),
  `Two-sided p-value` = c(
    chapter10_format_p(fisher_result$p.value),
    chapter10_format_p(midp_result$p.value),
    chapter10_format_p(barnard_result$p.value)
  )
```

```
)

smallsamplelab_apa_table(
  "10.2",
  "Comparison of exact p-values for the training example",
  exact_comparison_table,
  note = barnard_note,
  align = c("l", "l", "r")
)
```

Table 10.2

Comparison of exact p-values for the training example

Method	Conditioning	Two-sided p-value
Fisher's exact test	Conditions on both margins	0.070
Fisher's exact test with mid-p correction	Conditions on both margins, then applies a mid-p adjustment	0.038
Unconditional exact score test	Conditions on group sizes only	0.041

Note. The unconditional exact score test estimates a risk difference of 0.50, with a 95% confidence interval from 0.02 to 0.83.

For this table, Fisher's exact p-value is 0.070, whereas the mid-p correction gives 0.038 and the unconditional exact score test gives 0.041. The unconditional method is less conservative here because it does not condition on both margins. Reported as training minus no training, the estimated risk difference is 0.50, with a 95% confidence interval from 0.02 to 0.83.

Mid-p Corrections

Mid-p corrections reduce conservatism by subtracting half the probability of the observed table from the tail area. This yields a test that does not guarantee Type I error control at the nominal level in every configuration. Report mid-p results as a sensitivity analysis alongside the standard Fisher test rather than as a replacement default.

> ⚠ Common Misconception: "Exact" Means "Automatically Best"
>
> **Myth**: "If a test is exact, it must be the most appropriate or most informative option."
> **Reality**: Exactness describes how the p-value is computed under the null hypothesis. The appropriate conditioning scheme still depends on the design. In the training example above, Fisher's exact test gives 0.070, while the mid-p and unconditional exact alternatives give 0.038 and 0.041 respectively.

Lesson:

1. Exact computation and design appropriateness are separate questions.
2. Fisher's exact test is a strong default for confirmatory sparse 2×2 analyses.
3. Mid-p and unconditional exact tests are useful sensitivity checks when Fisher's conditioning may be overly restrictive.
4. With very small samples, the effect estimate and the underlying table can be more informative than a p-value alone.

Exact Binomial Test

The exact binomial test assesses whether the observed number of successes is compatible with a hypothesised success probability. It is appropriate for a single binary outcome when the null benchmark is known in advance or supplied by design.

Example: Exact Binomial Test

A clinic claims that 70% of patients improve with standard care. In a small audit of 15 patients, 13 improved. We test whether the observed proportion is consistent with the clinic's claim.

```
binom_result <- binom.test(x = 13, n = 15, p = 0.70, alternative =
 ↪  "two.sided")

binom_summary_table <- tibble(
  `Observed successes` = "13 of 15",
  `Observed proportion` = sprintf("%.3f", 13 / 15),
  `Hypothesised proportion` = sprintf("%.2f", 0.70),
  `95% exact CI` = sprintf("%.2f to %.2f", binom_result$conf.int[1],
   ↪  binom_result$conf.int[2]),
  `Exact p-value` = chapter10_format_p(binom_result$p.value)
)

smallsamplelab_apa_table(
  "10.3",
  "Exact binomial test summary",
  binom_summary_table,
  align = c("l", "r", "r", "l", "r")
)
```

Table 10.3

Exact binomial test summary

Observed successes	Observed proportion	Hypothesised proportion	95% exact CI	Exact p-value
13 of 15	0.867	0.70	0.60 to 0.98	0.258

The observed improvement proportion is 0.867, and the exact binomial p-value is 0.258. With such a small audit, the data remain compatible with the clinic's claimed 70% rate. The 95% exact confidence interval from 0.60 to 0.98 is wide enough to include both the claimed value and substantially higher improvement rates.

Exact Poisson Test

The exact Poisson test is used for count data when the quantity of interest is the number of events in a fixed amount of time, area, or exposure. It tests whether an observed count is consistent with a specified Poisson rate.

Example: Exact Poisson Test

A manufacturing process is expected to produce 3 defects per batch on average. In a random sample of 8 batches, we observe 32 total defects. We test whether the observed rate is consistent with the expected rate of 3 per batch.

```
poisson_observed_count <- 32
poisson_batches <- 8
poisson_expected_rate <- 3
poisson_result <- poisson.test(
  x = poisson_observed_count,
  T = poisson_batches,
  r = poisson_expected_rate,
  alternative = "two.sided"
)

poisson_summary_table <- tibble(
  `Observed count` = sprintf("%d defects across %d batches",
  ↪  poisson_observed_count, poisson_batches),
  `Observed rate` = sprintf("%.2f per batch", poisson_observed_count /
  ↪  poisson_batches),
  `Hypothesised rate` = sprintf("%.2f per batch", poisson_expected_rate),
  `95% exact CI` = sprintf("%.2f to %.2f", poisson_result$conf.int[1],
  ↪  poisson_result$conf.int[2]),
  `Exact p-value` = chapter10_format_p(poisson_result$p.value)
)

smallsamplelab_apa_table(
```

```
  "10.4",
  "Exact Poisson test summary",
  poisson_summary_table,
  align = c("l", "r", "l", "l", "r")
)
```

Table 10.4

Exact Poisson test summary

Observed count	Observed rate	Hypothesised rate	95% exact CI	Exact p-value
32 defects across 8 batches	4.00 per batch	3.00 per batch	2.74 to 5.65	0.102

The observed defect rate is 4.00 per batch, compared with the hypothesised rate of 3.00. The exact Poisson p-value is 0.102, so the data are broadly consistent with the expected defect rate. The 95% exact interval from 2.74 to 5.65 still includes 3.00.

Permutation Tests

Permutation tests compare the observed statistic with the distribution obtained by rearranging group labels under the null hypothesis. When every possible reallocation is enumerated, the resulting p-value is exact under the null. When only a large random sample of reallocations is used, the result is a Monte Carlo approximation. For two groups of size n1 and n2, the number of reallocations is (n1 + n2)! / (n1! n2!), equivalent to choosing n1 observations from the pooled sample. If this number is computationally feasible, for example below 100,000, enumerate all permutations; otherwise use a large Monte Carlo sample such as B = 10,000 and report both B and the random seed.

The key assumption is **exchangeability under the null hypothesis**: if the null is true, the joint distribution of the data remains the same under any reassignment of labels. That makes permutation tests especially attractive for small-sample comparisons where a parametric model would be difficult to justify.

Example: Permutation Test for Difference in Means

We compare test scores between two teaching methods with small groups of 8 students each. The outcome is continuous, but with such a small sample it is reasonable to avoid normality assumptions and to test the mean difference by permutation.

```r
method_a <- c(78, 82, 75, 88, 71, 85, 80, 83)
method_b <- c(68, 72, 76, 75, 79, 69, 73, 74)

all_scores <- c(method_a, method_b)
obs_diff <- mean(method_a) - mean(method_b)

perm_combos <- combn(seq_along(all_scores), length(method_a))
perm_diffs <- apply(
  perm_combos,
  2,
  function(idx) mean(all_scores[idx]) - mean(all_scores[-idx])
)
perm_p_value <- mean(abs(perm_diffs) >= abs(obs_diff))

perm_plot_data <- tibble(diff = perm_diffs)

ggplot(perm_plot_data, aes(x = diff)) +
  geom_histogram(
    binwidth = 1,
    boundary = 0,
    colour = "white",
    fill = "#A7D3F2"
  ) +
  geom_vline(xintercept = obs_diff, colour = "#C0392B", linewidth = 1) +
  geom_vline(xintercept = -obs_diff, colour = "#C0392B", linewidth = 1,
   ↪  linetype = "dashed") +
  labs(
    x = "Difference in means (A - B)",
    y = "Number of reallocations",
    title = "Exact permutation distribution of the mean difference",
    subtitle = "Red lines mark the observed difference\nand its mirror
     ↪  image under a two-sided test."
  ) +
  theme_classic(base_size = 12)
```

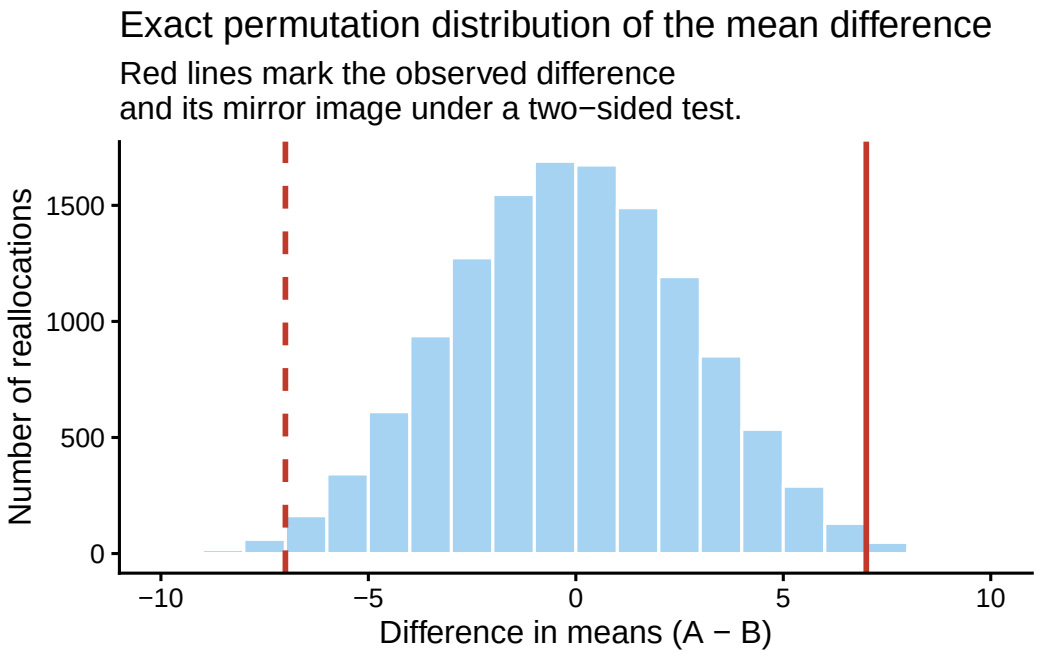


Figure 10.2: Exact permutation distribution for the mean difference between teaching methods.

Across all 12,870 possible reallocations, the observed mean difference is 7.00 points and the exact two-sided permutation p-value is 0.0126. Four-decimal precision is shown here because the exact enumeration produces a very small non-zero p-value. Because every possible label assignment is included, this result is exact rather than Monte Carlo. The figure makes the logic visible: the observed difference sits at the extreme tail of the permutation distribution.

Bootstrap Confidence Intervals

Bootstrap resampling constructs confidence intervals by repeatedly drawing samples with replacement from the observed data and recalculating the statistic of interest. The resulting bootstrap distribution approximates the sampling distribution of that statistic.

This chapter uses the **nonparametric bootstrap**, which resamples directly from the observed data. A **parametric bootstrap** simulates new datasets from a fitted parametric model, so its validity depends entirely on how well that model represents the data. For small-sample work, the nonparametric bootstrap is often the safer default when a model-free interval is desired.

Example: Bootstrap CI for the Median

We estimate the median recovery time for a small sample of patients and construct a 95% bootstrap confidence interval.

```
library(boot)

recovery_times <- c(12, 15, 14, 18, 16, 13, 17, 19, 14, 15, 20, 16, 15,
 ↪  18, 17)
sample_median <- median(recovery_times)

median_fun <- function(data, indices) {
  median(data[indices])
}

set.seed(2025)
boot_result <- boot(data = recovery_times, statistic = median_fun, R =
 ↪  2000)
boot_ci <- boot.ci(boot_result, conf = 0.95, type = c("perc", "bca"))

bootstrap_summary_table <- tibble(
  Statistic = "Median recovery time",
  Estimate = sprintf("%.0f days", sample_median),
  `95% percentile bootstrap CI` = sprintf(
    "%.0f to %.0f days",
    boot_ci$percent[4],
    boot_ci$percent[5]
  ),
  `95% BCa bootstrap CI` = sprintf(
    "%.0f to %.0f days",
    boot_ci$bca[4],
    boot_ci$bca[5]
  ),
  Resamples = format(boot_result$R, big.mark = ",")
)

smallsamplelab_apa_table(
  "10.5",
  "Bootstrap summary for median recovery time",
  bootstrap_summary_table,
  align = c("l", "r", "l", "l", "r")
)
```

Table 10.5

Bootstrap summary for median recovery time

Statistic	Estimate	95% percentile bootstrap CI	95% BCa bootstrap CI	Resamples
Median recovery time	16 days	14 to 17 days	14 to 17 days	2,000

The sample median recovery time is 16 days. The 95% percentile bootstrap interval runs from 14 to 17 days, and the BCa interval runs from 14 to 17 days. Both intervals summarise uncertainty around the median without requiring a normal-theory standard error formula. Percentile intervals are simple; BCa intervals adjust for bias and acceleration in the bootstrap distribution (Efron and Tibshirani 1993). With n < 30, BCa intervals can be unstable or fail when the statistic has many ties, so report the interval type and inspect the bootstrap distribution rather than treating BCa as automatically superior.

Key Takeaways

Exact tests are most valuable when the data are sparse enough that large-sample approximations become questionable. Fisher's exact test is a strong default for sparse 2×2 tables, but mid-p and unconditional exact procedures can serve as useful sensitivity checks when Fisher's conditioning is stronger than the design requires. Exact binomial and exact Poisson tests extend the same logic to single proportions and sparse event rates, while permutation tests rely on exchangeability and can be exact when every possible reallocation is enumerated. Bootstrap intervals are especially useful for statistics such as medians that lack simple analytic standard errors, provided the resampling procedure is reported transparently.

Self-Assessment Quiz

Test your understanding of exact tests and resampling methods from Chapter 10.

Question 1

When is Fisher's exact test especially appropriate?

a) When every expected cell count is comfortably above 20
b) When a 2×2 table has sparse counts and exact inference is preferred
c) When the outcome is continuous and normally distributed
d) When the goal is estimating a time-series trend

Question 2

Why can Fisher's exact test be conservative in some settings?

a) It ignores the observed table
b) It conditions on both margins, which can make rejection harder than necessary when both are not fixed by design
c) It always assumes normality
d) It uses too many bootstrap resamples

Question 3

When might Barnard-style unconditional tests or a mid-p correction be considered?

a) When you need a less conservative sensitivity check for sparse 2×2 data
b) When the outcome is continuous and homoscedastic
c) When all expected counts exceed 100
d) When fitting Bayesian regression with weakly informative priors

Question 4

What does the exact binomial test evaluate?

a) Whether two means differ after random permutation
b) Whether an observed number of successes is consistent with a hypothesised proportion
c) Whether a count process is overdispersed relative to Poisson
d) Whether a regression model has converged

Question 5

What is the exact Poisson test used for in this chapter?

a) Comparing two continuous variables with tied ranks
b) Testing whether an observed event count or rate is consistent with a specified Poisson rate
c) Selecting variables in penalised regression
d) Checking MCMC convergence

Question 6

What key assumption makes a permutation test valid under the null hypothesis?

a) Exchangeability of observations or labels under the null
b) Normal residuals with equal variance
c) At least 100 observations per group
d) A Poisson-distributed outcome

Question 7

Why is the bootstrap especially useful in small-sample work?

a) It guarantees narrow confidence intervals
b) It can approximate uncertainty for statistics like medians when closed-form standard errors are unavailable
c) It removes the need to report uncertainty
d) It is only valid for normal data

Question 8

Which reporting detail is most important to include for resampling procedures?

a) Only whether the result was significant
b) The number of resamples or permutations, the statistic used, and enough detail to reproduce the procedure
c) Only the sample size
d) Only the largest resampled statistic

Answers and Explanations

Question 1

Answer: b)

Explanation: Fisher's exact test is designed for sparse 2×2 tables where exact inference is preferred to the chi-square approximation. The chapter introduces it as the standard conditional test when some expected cell counts are small.

Question 2

Answer: b)

Explanation: The chapter explains that Fisher's test conditions on both margins. When the design leaves one margin free, that conditioning can make the test more conservative and reduce power.

Question 3

Answer: a)

Explanation: The chapter presents mid-p and unconditional exact procedures as less conservative alternatives or sensitivity checks for sparse 2×2 tables, especially when Fisher's conditioning may be overly restrictive.

Question 4

Answer: b)

Explanation: The exact binomial test is for a single-sample binary outcome. It compares the observed number of successes with a fixed benchmark proportion without relying on large-sample approximations.

Question 5

Answer: b)

Explanation: The chapter uses the exact Poisson test for sparse count data when the interest is whether an observed count or event rate differs from a known benchmark.

Question 6

Answer: a)

Explanation: Permutation tests rely on exchangeability under the null hypothesis. If the joint distribution of the data remains the same under any reassignment of labels, the permutation reference distribution is valid.

Question 7

Answer: b)

Explanation: The bootstrap is useful when the statistic of interest does not have a simple analytic standard error or when a model-free interval is preferred. The chapter's median example illustrates this directly.

Question 8

Answer: b)

Explanation: The chapter states that resampling analyses should report the statistic being resampled, the number of permutations or bootstrap resamples, the random seed, and whether the result is exact or Monte Carlo.

Chapter 11: Nonparametric Rank-Based Methods

Learning Objectives

By the end of this chapter, you will be able to explain why rank-based methods are useful for ordinal, skewed, or outlier-prone outcomes, choose among Mann–Whitney, Wilcoxon signed-rank, Kruskal–Wallis, Friedman, Spearman, and Kendall procedures, interpret rank-based tests as location shifts only when distributional assumptions permit, and report p-values alongside robust effect sizes and uncertainty.

When Rank-Based Methods Help

Rank-based methods replace raw observations with their ranks. That makes them less sensitive to extreme values and less dependent on normality than mean-based methods. The trade-off is that ranks discard some information about magnitude, so a t-test or linear model may be more efficient when assumptions are reasonable and the outcome scale is genuinely interval.

In small-sample work, rank-based tests are most useful for ordinal outcomes, visibly skewed continuous outcomes, and settings where one or two extreme observations would dominate a mean. They are not magic assumption-free substitutes for thinking about the design. Independence, pairing, similar distributional shape, and the substantive meaning of ranks still matter.

Mann–Whitney U Test

The Mann–Whitney U test, also called the Wilcoxon rank-sum test, compares two independent groups by ranking all observations together and evaluating whether one group tends to receive larger ranks. This chapter uses **Mann–Whitney U** for the two-sample procedure and refers to the Wilcoxon rank-sum statistic only when describing the statistic returned by R. When the two groups have similar distributional shapes, including similar variance and skew, the result can be described as evidence about a median or location shift. If the shapes differ markedly, the safer interpretation is stochastic dominance: a randomly selected observation from one group tends to be larger than a randomly selected observation from the other. Inspect histograms, density plots, or dotplots before choosing the interpretation.

Interpreting Mann–Whitney when shapes differ

Use this decision rule before writing the result. First inspect histograms, dotplots, or empirical cumulative distribution functions. If the two groups have broadly similar shape and spread, a location-shift interpretation is reasonable. If the shapes differ, do not describe the result as a simple median comparison. Report stochastic dominance using a probability-of-superiority measure such as Cliff's delta, together with medians and IQRs for context.

Figure 11.1 shows why this distinction matters. The two groups have the same median, but the spread group has more observations in both tails. A rank test can be sensitive to this broader distributional difference, so the interpretation should be about relative ranks or stochastic dominance rather than a median shift.

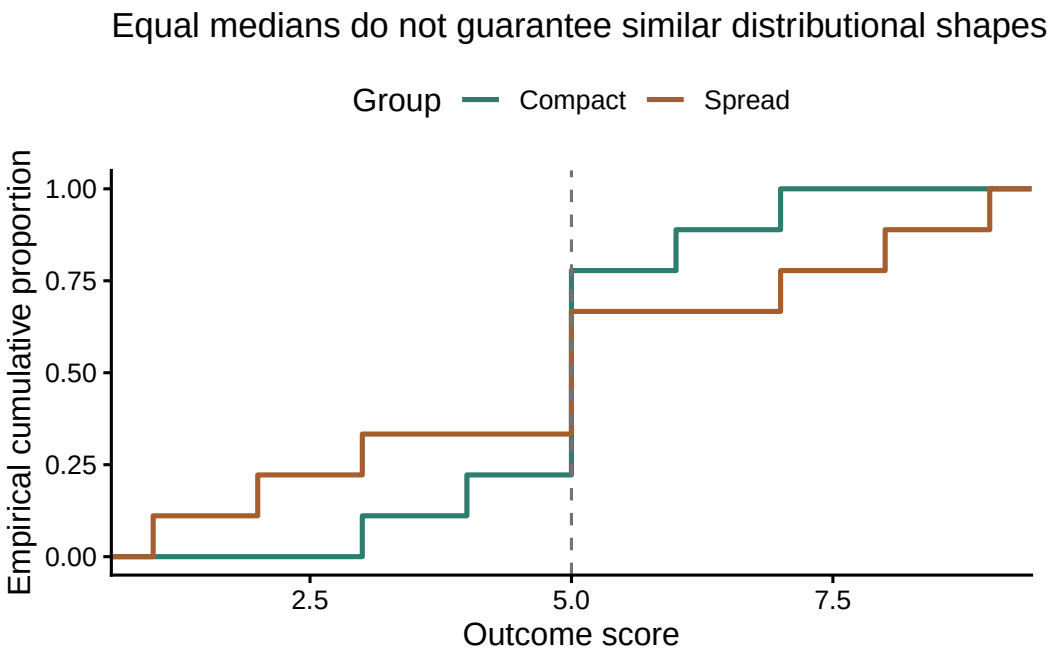


Figure 11.1: Empirical cumulative distributions with identical medians but different shapes.

Table 11.1

Equal medians with different distributional shapes

group	Median	IQR	Minimum	Maximum
Compact	5	0	3	7
Spread	5	4	1	9

Note. Both groups have median = 5. The spread group has a wider distribution, so a rank-test

result would need a stochastic-dominance interpretation rather than a median-difference interpretation.

In the wait-time example, Branch A has shorter waits than Branch B. Table 11.2 gives the descriptive context, and Table 11.3 gives the inferential summary.

Table 11.2

Wait-time descriptives by branch

Branch	n	Median	IQR	Minimum	Maximum
A	10	7.5	2.5	5	12
B	12	13.0	2.5	10	16

Note. Wait time is measured in minutes. The distributions should be inspected before interpreting the rank-sum test as a simple median comparison.

Table 11.3

Mann–Whitney test and Cliff's delta for the wait-time example

Test	W statistic	Hodges–Lehmann shift (A - B)	95% CI	p-value	Cliff's delta (A vs B)	Delta 95% CI
Mann–Whitney U	4.5	-5.0 minutes	-7.0 to -3.0	< 0.001	-0.925	-1.00 to -0.73

Note. The negative shift and negative Cliff's delta indicate that Branch A wait times tend to be lower than Branch B wait times. The bootstrap CI reflects effect-size uncertainty in this small sample.

The evidence is strong that the wait-time distributions differ. The estimated location shift is -5.0 minutes for Branch A minus Branch B, so Branch A tends to have shorter waits. Cliff's delta is -0.925 with a bootstrap 95% CI from -1.00 to -0.73, meaning that a randomly selected Branch A wait is usually lower than a randomly selected Branch B wait but the exact magnitude remains sample-dependent.

Effect Sizes Can Be Unstable in Tiny Samples

Large effect estimates in small samples can arise from real separation, but they can also arise from ordinary sampling variation. Table 11.4 illustrates this with two groups generated from the same normal distribution. The observed Cohen's d carries meaning only when read alongside the sample size, p-value, confidence interval, and substantive plausibility.

Table 11.4

A large-looking effect from two identical populations

Quantity	Value
Group A mean	51.9
Group B mean	49.6
Observed Cohen's d	0.24
Welch t-test p-value	0.719

Note. Both groups were generated from the same population with mean 50 and standard deviation 10. The example shows why effect sizes from n = 5 per group should be treated as provisional.

Wilcoxon Signed-Rank Test

The Wilcoxon signed-rank test is the paired-sample counterpart to the Mann–Whitney test. It ranks the absolute paired differences and tests whether positive and negative ranks balance around zero. The usual location-shift interpretation assumes that the distribution of paired differences is roughly symmetric. The pseudomedian estimate equals the median paired difference only under that symmetry. With skewed paired differences, report the pseudomedian and confidence interval without calling it the median.

The pseudomedian is the median of all Walsh averages: each paired difference is averaged with itself and with every other paired difference, giving $n(n+1)/2$ values. In this example, 12 paired differences produce 78 Walsh averages. That definition explains why the signed-rank estimate can differ from the ordinary sample median when the paired-difference distribution is skewed.

In the intervention example, anxiety scores decline after treatment. Table 11.5 shows the paired summary.

Table 11.5

Wilcoxon signed-rank summary for paired anxiety scores

Median before	Median after	Median improvement	V statistic	Pseudomedian shift	95% CI	p-value
70	65	5	78	5.0	4.5 to 5.5	0.002

Note. Differences are coded as before minus after, so positive values indicate improvement.

The signed-rank test gives V = 78 and p = 0.002. The estimated pseudomedian improvement is about 5.0 points, with a confidence interval that excludes zero.

Kruskal–Wallis and Friedman Tests

Kruskal–Wallis extends the rank-sum idea to three or more independent groups. A significant result says that at least one group distribution differs, but it does not identify the pair responsible. Follow-up pairwise comparisons require a multiplicity adjustment.

The workflow is the same each time: rank all observations across groups, compute the omnibus Kruskal–Wallis statistic, estimate an effect size such as epsilon-squared, and then run adjusted pairwise comparisons only if the omnibus result is worth following up. In the ward-satisfaction example, the mean ranks show the direction of the pattern before the test is interpreted.

```
satisfaction_data %>%
  mutate(rank = rank(score, ties.method = "average")) %>%
  group_by(ward) %>%
  summarise(mean_rank = mean(rank), .groups = "drop")

kruskal.test(score ~ ward, data = satisfaction_data)

rstatix::dunn_test(
  satisfaction_data,
  score ~ ward,
  p.adjust.method = "holm"
)
```

The epsilon-squared estimate is computed as $(H - k + 1)/(n - k)$, where H is the Kruskal–Wallis statistic, k is the number of groups, and n is the total sample size (Tomczak and Tomczak 2014). It is a descriptive measure of how strongly ranks differ across groups, not a replacement for the design context.

Table 11.6

Kruskal–Wallis and adjusted pairwise comparisons

Result	Statistic	df	p-value	Effect
Kruskal–Wallis	12.25	2	0.002	0.68
Blue vs Green			0.001	
Blue vs Red			0.102	
Green vs Red			0.124	

Note. Pairwise rows report Holm-adjusted Dunn test p-values.

Friedman's test handles three or more related conditions. For the task-condition example, Table 11.7 reports a large within-person rank effect.

Friedman's test ranks conditions within each participant rather than ranking all observations together. Kendall's W rescales the Friedman statistic to an effect-size measure from 0 to 1, where larger values indicate stronger separation among the repeated conditions. If the omnibus test is followed up, paired rank comparisons should again be adjusted for multiplicity.

```
friedman.test(score ~ condition | participant, data = performance_long)

pairwise.wilcox.test(
  performance_long$score,
  performance_long$condition,
  paired = TRUE,
  p.adjust.method = "holm",
  exact = FALSE
)
```

Table 11.7

Friedman test summary for repeated task conditions

Test	Chi-square	df	p-value	Kendall's W
Friedman	12.07	2	0.002	0.75

Note. Kendall's W is an effect-size measure for agreement or separation among repeated-measure ranks.

Table 11.8

Adjusted paired comparisons after the Friedman test

Comparison	Holm-adjusted p
condition_2 vs condition_1	0.025
condition_3 vs condition_1	0.040
condition_3 vs condition_2	0.240

Note. Pairwise rows report Holm-adjusted paired Wilcoxon p-values. These comparisons are descriptive follow-ups to the omnibus Friedman result.

Rank Correlations

Spearman's rho and Kendall's tau measure monotonic association without requiring a linear relationship or normally distributed variables. Spearman's rho is the Pearson correlation of ranks. Kendall's tau is based on concordant and discordant pairs, so it is often easier to interpret when tied ranks are common.

When there are no tied ranks, `cor.test()` can compute exact small-sample p-values for Spearman or Kendall by setting `exact = TRUE`. When ties are present, R uses approximate p-values. If exact inference is important with ties, use a permutation procedure and report it explicitly.

```
cor.test(x, y, method = "spearman", exact = TRUE)
cor.test(x, y, method = "kendall", exact = TRUE)
```

Table 11.9

Rank-correlation summaries for experience and satisfaction

Statistic	Estimate	p-value	Interpretation
Spearman's rho	0.962	< 0.001	Strong monotonic association
Kendall's tau	0.911	< 0.001	Strong concordance between ranks

Note. Both tests use approximate p-values because the data contain ties. If exact small-sample p-values matter, use a permutation procedure or specialised software and report the method.

Lab Practical 11.1: Sales Performance Analysis

A retail company piloted two training programmes across 10 stores each and collected customer satisfaction scores on a 1 to 100 scale. The question is whether Training B produces higher satisfaction than Training A. Figure 11.2 shows nearly complete separation between the programmes, and Table 11.10 reports the descriptive and inferential summaries.

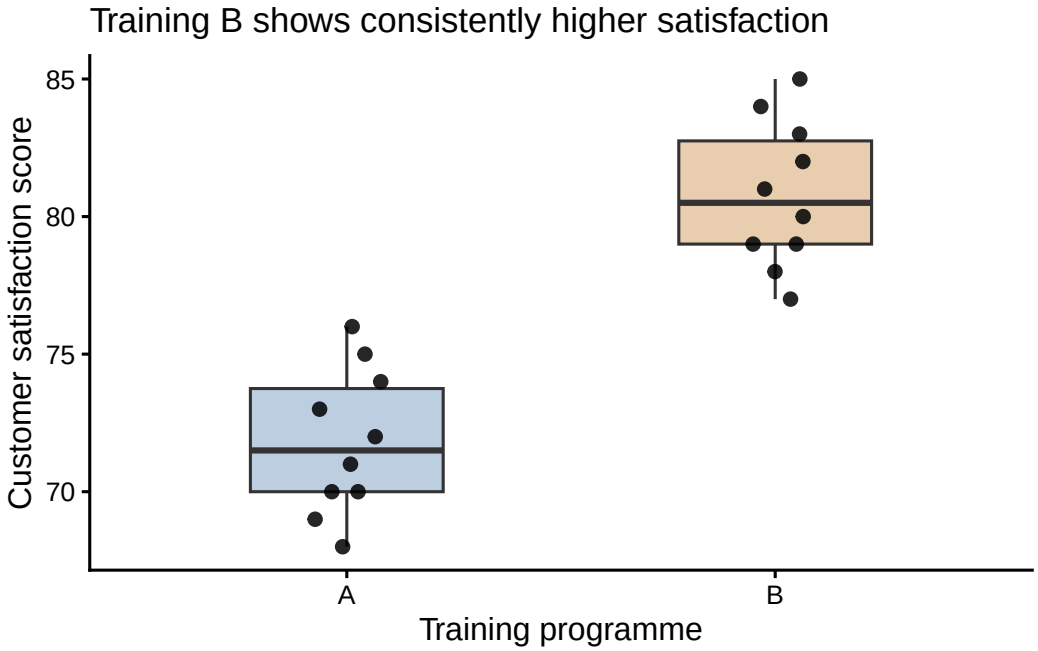


Figure 11.2: Customer satisfaction scores by training programme.

Table 11.10

Sales-training descriptives and rank-sum test

Measure	Value	Details
Programme A	n = 10; median = 71.5; IQR = 3.8	Range 68 to 76
Programme B	n = 10; median = 80.5; IQR = 3.8	Range 77 to 85
Mann–Whitney U	W = 0; p < 0.001	Two-sided rank-sum test
Hodges–Lehmann shift (A - B)	-9 points	95% CI: -12 to -6
Cliff's delta (A vs B)	-1.00	Bootstrap 95% CI: -1.00 to -1.00

Note. The negative shift and Cliff's delta occur because the comparison is coded as A minus B. Substantively, Training B is higher. The bootstrap delta interval is degenerate here because all observed Training B scores exceed all observed Training A scores.

The corrected result is stronger than the older draft suggested: W = 0, p < 0.001, and Cliff's delta is -1.00 with a bootstrap 95% CI from -1.00 to -1.00. Because all Training B scores exceed all

Training A scores, this is complete stochastic separation in the sample. The reporting language should still acknowledge the pilot design rather than claiming guaranteed future superiority.

Choosing Among Rank-Based Methods

The methods in this chapter differ by design, not by which one seems most familiar. Start from the study structure: independent groups, paired observations, repeated conditions, or association between two ordered variables. Then report an effect size that matches the design rather than relying on the p-value alone.

Table 11.11

Rank-based method selection guide

Question	Test	Effect size	R function
Two independent groups	Mann–Whitney U	Cliff's delta or rank-biserial correlation	wilcox.test(); rstatix::wilcox_effsize()
Two paired measurements	Wilcoxon signed-rank	Rank-biserial correlation; pseudomedian shift with CI	wilcox.test(paired = TRUE)
Three or more independent groups	Kruskal–Wallis	Epsilon-squared plus adjusted pairwise contrasts	kruskal.test(); rstatix::dunn_test()
Three or more repeated conditions	Friedman	Kendall's W plus adjusted paired contrasts	friedman.test(); pairwise.wilcox.test(paired = TRUE)
Monotonic association	Spearman or Kendall rank correlation	rho or tau with exact or permutation p-value when feasible	cor.test(method = "spearman" or "kendall")

Note. Use visual checks and design knowledge before choosing the interpretation. A rank test is not automatically a median test when distributional shapes differ.

Reporting Rank-Based Results

A rank-based result should not be reported as a p-value alone. The reader needs to know the design, the outcome scale, the group summaries, the test statistic, the effect size, and the interpretation chosen. For independent groups, that usually means reporting medians and IQRs by group, the Mann–Whitney U result, the p-value, and a probability-style effect size such as Cliff's delta or rank-biserial correlation. For paired designs, report the median paired change, the signed-rank statistic, the Hodges–Lehmann pseudomedian shift and its interval where available. For more than two groups, state whether follow-up comparisons were adjusted for multiplicity.

Use median-shift language only when the distributions have broadly similar shapes. If one group is more variable or more skewed, write the conclusion as stochastic dominance: observations from one group tended to be higher than observations from the other. This distinction is not cosmetic. It tells readers whether the analysis is about a typical location shift or about the ordering of observations across the full distributions.

A concise report might read: "Satisfaction scores were higher in Training B than Training A. Because the group shapes were similar, the rank-sum result was interpreted as a location shift, W = 0, p < .001, Hodges–Lehmann shift = -8.5 points, Cliff's delta = -1.00. The negative sign reflects the A-minus-B coding. Substantively, all observed Training B scores exceeded the corresponding Training A range." If the shapes had differed, the same result should be framed as evidence that Training B scores tended to be higher, not as proof of a median difference.

Key Takeaways

Rank-based tests are useful when outcomes are ordinal, skewed, tied, or vulnerable to outliers, but they still require attention to design and interpretation. Mann–Whitney and Wilcoxon signed-rank tests address independent and paired two-sample questions. Kruskal–Wallis and Friedman extend rank comparisons to multiple independent or repeated groups. Spearman's rho and Kendall's tau summarise monotonic association. In small samples, rank-based p-values should be reported with medians, IQRs, robust effect sizes, and enough context to distinguish a defensible pattern from sampling noise. Robust mean-based alternatives are also worth considering when the outcome scale remains meaningfully continuous (Mair and Wilcox 2020).

Self-Assessment Quiz

Test your understanding of nonparametric tests and rank-based methods from Chapter 11.

Question 1

The Mann–Whitney U test is most appropriate when:

a) The outcome is a normally distributed continuous variable and means are the only estimand
b) Two independent groups have ordinal or skewed continuous outcomes
c) The same participants are measured under three conditions
d) The goal is to estimate a Poisson event rate

Question 2

A significant Mann–Whitney test can be interpreted as a median difference only when:

a) Both groups have similar distributional shapes
b) The p-value is below 0.001
c) The sample size is larger than 200
d) There are no repeated values

Question 3

The Wilcoxon signed-rank test is used for:

a) Two independent groups
b) Paired or matched observations
c) Three or more independent groups
d) A single count against a benchmark rate

Question 4

What does the Hodges–Lehmann estimate summarise in a rank-sum comparison?

a) The ordinary mean difference
b) The median of all pairwise differences
c) The chi-square statistic
d) The number of tied ranks

Question 5

After a significant Kruskal–Wallis test, the next step is usually to:

a) Assume every pair differs
b) Use adjusted pairwise rank comparisons to identify where the difference lies
c) Switch to Pearson correlation
d) Ignore the result because it is nonparametric

Question 6

Friedman's test is the rank-based alternative to:

a) Repeated-measures ANOVA
b) One-sample t-test
c) Poisson regression
d) Fisher's exact test

Question 7

Kendall's tau is often useful when:

a) The outcome is a sparse event count
b) There are many tied ranks or a probability-based rank interpretation is helpful
c) The model has many predictors
d) A bootstrap median interval is required

Question 8

A large Cliff's delta from a tiny sample should be interpreted:

a) As definitive proof of a population effect
b) Together with the sample size, confidence interval, p-value, and substantive plausibility
c) As invalid because rank tests cannot have effect sizes
d) As equivalent to a mean difference in standard deviations

Answers and Explanations

Question 1

Answer: b)

Explanation: Mann–Whitney is the independent two-group rank test. It is especially useful for ordinal outcomes or continuous outcomes where skewness or outliers make mean-based inference fragile.

Question 2

Answer: a)

Explanation: The median-shift interpretation depends on similar distributional shapes, including similar variance and skew. If the shapes differ, stochastic dominance is the safer interpretation.

Question 3

Answer: b)

Explanation: The signed-rank test is the nonparametric paired-sample procedure. It ranks paired differences and tests whether they are centered around zero.

Question 4

Answer: b)

Explanation: For a two-sample rank-sum comparison, the Hodges–Lehmann estimate is the median of all pairwise differences and is interpreted as a robust location shift.

Question 5

Answer: b)

Explanation: The omnibus Kruskal–Wallis test says at least one distribution differs. Pairwise comparisons with multiplicity correction are needed to identify the specific contrasts.

Question 6

Answer: a)

Explanation: Friedman's test compares three or more related or repeated-measure conditions using within-participant ranks.

Question 7

Answer: b)

Explanation: Kendall's tau is based on concordant and discordant pairs, making it useful when tied ranks are common and when that probability interpretation helps readers.

Question 8

Answer: b)

Explanation: With small samples, large effect estimates can be real or can arise by chance. The effect estimate needs uncertainty, context, and preferably replication.

Chapter 12: Methods for Sparse Counts and Short Time Series

Learning Objectives

By the end of this chapter, you will be able to recognise when sparse counts and very short time series make standard large-sample modelling fragile, use exact Poisson procedures for small event counts, compare event rates with appropriate uncertainty, construct transparent forecast intervals for short series, and report overdispersion or zero-inflation without overfitting models the data cannot support.

The Challenge of Sparse Counts

Sparse counts occur when events are rare, exposure is limited, or most units have zero events. In those settings, ordinary Poisson regression can produce unstable standard errors, asymptotic tests can be inaccurate, and zero-inflated or negative binomial models may be too parameter-heavy for the available data. The small-sample goal is usually not to build an elaborate model, but to make a defensible comparison, quantify uncertainty, and explain what the data can and cannot support.

Short time series create a parallel problem. With fewer than about 30 observations, classical ARIMA identification is usually underpowered: there are too few data points to estimate auto-correlation structure reliably. Simple trend summaries, residual-bootstrap intervals, and clearly labelled descriptive forecasts are often more honest than a complex model that appears precise only because it is overfit.

Choosing a Sparse-Data Method

Begin with the data structure rather than the model name. A sparse-count analysis is usually trying to answer one of four questions: whether a count differs from a benchmark rate, whether two rates differ, whether count variability is larger than a Poisson model expects, or whether a short sequence supports a cautious forecast. Each question has a different minimum reporting unit.

Situation	Useful starting point	What must be reported
One small count compared with a known exposure-adjusted benchmark	Exact Poisson test	Event count, exposure, benchmark rate, exact CI and two-sided p-value
Two independent sparse event rates	Exact Poisson rate-ratio comparison	Counts and exposures in both groups, rate ratio, exact CI and p-value
Count outcome with variance much larger than the mean	Poisson summary plus quasi-Poisson sensitivity check	Mean, variance, dispersion estimate, coefficient scale and adjusted standard errors
Overdispersed counts with enough information for an extra parameter	Negative binomial sensitivity analysis	Sample size, theta estimate, standard errors and whether the fit is stable
Many zeros in a very small dataset	Descriptive zero-count reporting before any zero-inflated model	Number and proportion of zeros, total events and why a ZIP model is or is not defensible
Short series with fewer than about 15 observations	Descriptive trend and exploratory interval only	Number of time points, trend assumption, resampling unit, seed and bootstrap resamples

This sequence prevents a common small-sample error: fitting the most specialised model first. Exact and descriptive methods usually give the clearest account of the evidence. More complex models are useful only when they answer a question that the simpler summaries cannot answer and when the data contain enough information to estimate the extra parameters.

Exact Poisson Test for a Benchmark Rate

The exact Poisson test compares an observed count with a pre-specified event rate over a known exposure. In the quality-control example, 15 defects were observed across five batches, against a benchmark of two defects per batch. Table 12.1 shows that the observed rate is higher, but the exact interval still includes the target rate.

Table 12.1

Exact Poisson test for the defect-rate benchmark

Quantity	Value
Observed defects	15
Exposure (batches)	5 batches
Observed rate	3.00 defects per batch

Quantity	Value
Target rate	2.00 defects per batch
Exact p-value	0.113
95% exact CI for rate	1.68 to 4.95

Note. The exact test compares 15 observed defects over five batches with a target rate of two defects per batch.

The exact two-sided p-value is 0.113, so this small trial does not provide strong evidence that the process rate differs from the benchmark of 2.00 defects per batch at alpha = 0.05. The practical warning remains meaningful: the point estimate is 3.00 defects per batch, and the 95% exact CI (1.68 to 4.95) is wide because only 15 events were observed.

Comparing Two Sparse Event Rates

For two independent event counts, compare rates using the event counts and the exposure times together. Table 12.2 shows the raw rates for two trials, and Table 12.3 reports the exact Poisson rate-ratio comparison.

Table 12.2

Adverse event rates by trial

Trial	Events	Patient-days	Rate per patient-day
A	8	50	0.160
B	3	45	0.067

Note. Rates are events divided by patient-days.

Table 12.3

Exact comparison of adverse event rates

Quantity	Value
Rate ratio (A/B)	2.40
95% exact CI	0.58 to 14.05
Exact p-value	0.234

Note. The exact rate-ratio CI is computed on the log scale and back-transformed. With only 11 total events, the log-scale standard error is large, producing a wide interval on the original rate-ratio scale.

Trial A's observed rate is higher, with an estimated rate ratio of 2.40. The 95% exact CI (0.58 to 14.05) includes 1 and spans more than an order of magnitude, so the result should be framed as imprecise rather than as evidence of a clear difference. Report the raw event counts and exposure times alongside the ratio so readers can assess how much information supports the estimate.

Bootstrap Forecast Intervals for Short Time Series

With 12 monthly observations, a simple linear trend can be easier to explain than a fitted ARIMA model. Figure 12.1 shows the observed series, the fitted linear trend, and a residual-bootstrap interval for month 13 (Efron and Tibshirani 1993). Table 12.4 gives the numeric summary.

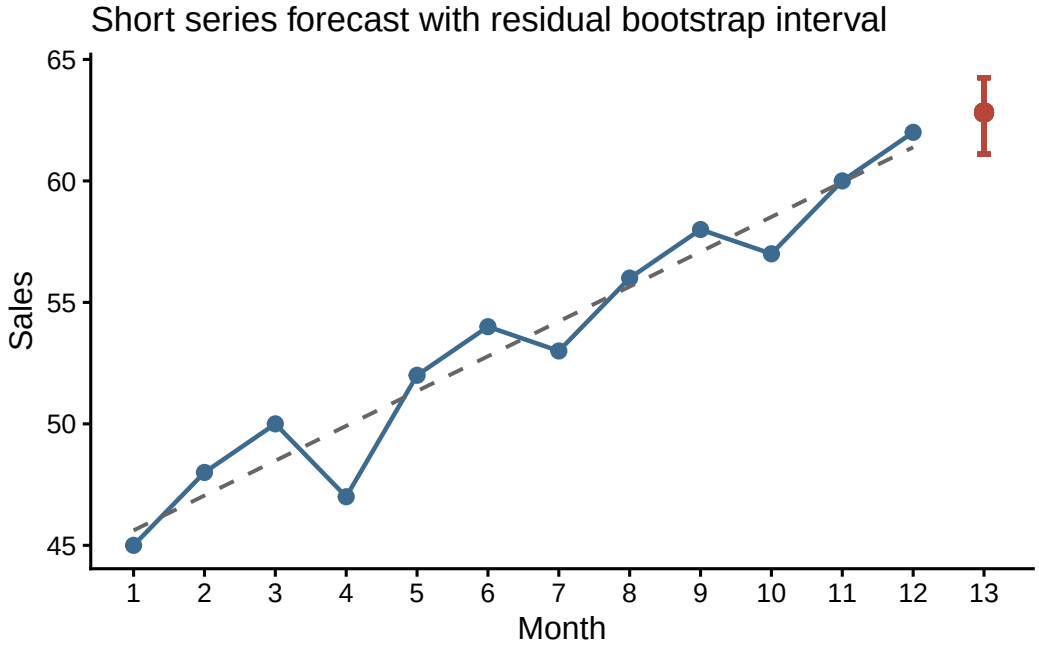


Figure 12.1: Short monthly sales series with a residual-bootstrap forecast interval for month 13. The interval reflects residual variability around the fitted linear trend but not model-selection or autocorrelation uncertainty.

Table 12.4

Short-series forecast summary

Quantity	Value
Observed months	12
Trend slope	1.43 sales units per month
Month 13 point forecast	62.8
95% residual-bootstrap interval	61.1 to 64.3

Quantity	Value
Residual autocorrelation check	Ljung-Box p = 0.008

Note. This residual bootstrap resamples model residuals with replacement, adds them to fitted values, refits the trend in each bootstrap sample, and forecasts month 13. It assumes residuals are approximately independent and identically distributed; autocorrelation or structural breaks would require a block bootstrap or state-space approach. The Ljung-Box row is a small-sample screen, not proof of independence.

The forecast is transparent, but it should not be oversold. With only 12 observations, the apparent trend may be sensitive to a few months, and the residual-bootstrap interval is an exploratory forecast band rather than a definitive prediction interval. If the residuals show visible autocorrelation or a structural break, the resampling unit is wrong and this interval should not be used without a time-series-specific bootstrap.

Zero Inflation and Overdispersion

Zero-inflated models estimate both a count process and an extra-zero process, typically adding at least two parameters relative to a standard Poisson model. That can be useful with enough data, such as n = 50 to 100 with clear zero inflation, but in very small samples (n < 30) it often creates more parameters than the dataset can support, leading to unstable estimates or non-convergence (Cameron and Trivedi 2013). A pragmatic first step is to report the proportion of zeros, the mean, and the variance, then decide whether a simple exact or quasi-likelihood approach is adequate.

Table 12.5 summarises the simulated clinic-visit counts. The variance exceeds the mean in both groups, which is a warning sign for overdispersion.

Table 12.5

Clinic visit count summaries by treatment group

Treatment	n	Mean	Variance	Variance/mean	Zero-count cases
Standard	6	2.83	5.37	1.89	1
Enhanced	6	4.83	6.97	1.44	0

Note. When the variance is materially larger than the mean, for example a variance/mean ratio above about 1.5, a standard Poisson model may understate uncertainty. Quasi-Poisson or negative binomial models can provide more realistic standard errors, but with very small samples the dispersion estimate is itself uncertain.

For a fitted Poisson model, the same issue can be checked directly with `performance::check_overdispersion()`. Use this as a diagnostic prompt rather than a mechanical decision rule; with very small samples, the dispersion estimate can move substantially when one observation changes.

```
poisson_fit <- glm(visits ~ treatment, family = poisson(), data =
  ↪  visit_data)
performance::check_overdispersion(poisson_fit)
```

Quasi-Poisson as a Small-Sample Adjustment

A quasi-Poisson model keeps the Poisson mean structure but estimates a dispersion parameter, phi, and inflates standard errors by sqrt(phi). The coefficient estimates remain identical to those from the standard Poisson model; only the standard errors and derived p-values change. When phi > 1, the quasi-Poisson result is more conservative; when phi is close to 1, the two models converge.

Table 12.6

Poisson and quasi-Poisson comparison for clinic visit counts

Model	Log rate ratio	Std. error	p-value	Dispersion estimate
Poisson	0.53	0.31	0.080	1.00 (fixed)
Quasi-Poisson	0.53	0.39	0.206	1.67

Note. The coefficient compares Enhanced with Standard treatment on the log-rate scale. The quasi-Poisson log rate ratio is 0.53, corresponding to a rate ratio of 1.71. The dispersion parameter phi is estimated as the Pearson chi-square statistic divided by residual degrees of freedom; here phi = 1.67, inflating standard errors by about 1.29. p-values are from the Wald test on the log scale. The displayed SE values are rounded; the printed phi is computed from the full Pearson statistic and may not reproduce exactly from rounded SEs.

In this example, the standard Poisson model would make the treatment contrast look more precise than the data justify. The quasi-Poisson result is a useful sensitivity check, but with only 12 observations it should still be reported as exploratory.

Negative Binomial and Zero-Inflated Alternatives

Negative binomial regression is another common response to overdispersed count data. Unlike quasi-Poisson, it specifies a mean-variance relationship and estimates an additional dispersion parameter, usually called theta. That extra parameter can be useful when there is enough information, but with very small samples it is itself unstable. Table 12.7 therefore treats the negative binomial model as a sensitivity check rather than a replacement for the simpler summaries above.

Table 12.7

Negative binomial sensitivity check for clinic visit counts

Quantity	Value
Log rate ratio	0.53
Rate ratio	1.71
Std. error	0.37
p-value	0.146
Theta	7.94

Note. The negative binomial model is fitted with MASS::glm.nb(). With n = 12, the theta estimate is included only as a sensitivity check; the quasi-Poisson and raw count summaries remain central.

The same caution applies more strongly to zero-inflated Poisson models. A zero-inflated model would estimate both the visit-count process and an extra-zero process. In this example there are only 12 observations and only one zero-count case, so a zero-inflated model would add parameters without enough information to estimate them reliably. The defensible report is therefore descriptive: show the zero count, the mean-variance pattern, and the Poisson/quasi-Poisson sensitivity comparison rather than fitting a model that the data cannot support.

Reporting Checklist for Sparse-Count Analyses

State the observed event count and exposure, such as "15 defects across 5 batches." Report the exact method used, the point estimate, the 95% exact CI, and the two-sided p-value. For rate ratios, give raw counts and exposures for both groups. For resampling procedures, report the random seed and number of bootstrap resamples, such as `set.seed(2025)` and `R = 2000`. If using quasi-Poisson, report the dispersion estimate and note that coefficients are unchanged from the Poisson model while standard errors are adjusted. Frame results as exploratory when n < 30 or total events < 20.

Key Takeaways

Sparse counts and short time series reward restraint. Exact Poisson tests provide defensible inference for small event counts and benchmark comparisons, while exact rate-ratio intervals show how quickly uncertainty grows when total events are few. For short series, simple trend models and residual-bootstrap intervals are often more transparent than complex ARIMA models when their assumptions are stated (Hyndman and Athanasopoulos 2021). Zero-inflated and negative binomial models can be appropriate with enough information, but with very small samples researchers should begin with descriptive summaries, dispersion checks, exact tests, and clearly qualified uncertainty.

Self-Assessment Quiz

Question 1

When is an exact Poisson test most useful?

a) For normally distributed continuous outcomes
b) For a small observed count compared with a known exposure-adjusted benchmark rate
c) For estimating a factor model
d) For checking whether two medians differ

Question 2

In the defect example, 15 defects observed across five batches gives what observed rate per batch?

a) 2 defects per batch
b) 3 defects per batch
c) 5 defects per batch
d) 15 defects per batch

Question 3

A rate-ratio confidence interval that includes 1 should usually be described as:

a) Clear evidence of different rates
b) Inconclusive or imprecise evidence for a rate difference
c) Proof that the rates are identical
d) A reason to ignore exposure time

Question 4

Why are ARIMA models risky with very short series?

a) They cannot use time as a variable
b) There are usually too few observations to estimate autocorrelation structure reliably
c) They require binary outcomes
d) They always ignore uncertainty

Question 5

What does overdispersion mean for count data?

a) The variance is larger than the mean
b) The mean is exactly zero
c) Every count is identical
d) The data are normally distributed

Question 6

What does the quasi-Poisson model change compared with standard Poisson regression?

a) It changes the outcome from counts to ranks
b) It estimates a dispersion parameter and adjusts standard errors
c) It removes all zero counts
d) It requires no exposure information

Question 7

Why are zero-inflated models difficult in very small samples?

a) They estimate both a zero process and a count process
b) They cannot represent zeros
c) They require normally distributed residuals
d) They are identical to exact Poisson tests

Question 8

A residual-bootstrap forecast interval for a 12-point time series should be reported as:

a) A definitive prediction interval with no caveats
b) A transparent exploratory interval that may not capture all model uncertainty
c) Invalid because bootstrap methods cannot use time series
d) Equivalent to a structural time-series model

Answers and Explanations

Question 1

Answer: b)

Explanation: The exact Poisson test is designed for event counts observed over a known exposure, especially when the count is small and large-sample approximations are questionable.

Question 2

Answer: b)

Explanation: The observed rate is the event count divided by exposure: 15 defects over five batches equals 3 defects per batch.

Question 3

Answer: b)

Explanation: A rate ratio of 1 represents equal rates. If the interval includes 1, the data do not provide clear evidence of a difference at the chosen confidence level.

Question 4

Answer: b)

Explanation: Short series contain too little information for stable identification and estimation of autoregressive and moving-average terms.

Question 5

Answer: a)

Explanation: A Poisson model assumes the variance equals the mean. Overdispersion occurs when the observed variance is larger, which can make ordinary Poisson standard errors too small and p-values anti-conservative. Quasi-Poisson or negative binomial models address this by estimating extra dispersion.

Question 6

Answer: b)

Explanation: Quasi-Poisson keeps the same mean model but uses an estimated dispersion parameter to inflate standard errors when variability exceeds the Poisson assumption.

Question 7

Answer: a)

Explanation: Zero-inflated models add parameters for the extra-zero process. With very small samples, those extra parameters can be unstable or unidentified.

Question 8

Answer: b)

Explanation: Residual bootstrapping is useful for capturing residual variability around a fitted trend, but with very short series it may not capture model-selection uncertainty, autocorrelation structure, or all coefficient uncertainty. It should be reported as an exploratory interval.

Chapter 13: Penalised and Bayesian Regression for Small Samples

Learning Objectives

By the end of this chapter, you will be able to explain why ordinary maximum likelihood can fail with sparse regression data, recognise separation in logistic regression, use penalised estimates to stabilise small-sample models, describe how weakly informative Bayesian priors regularise estimates, and report sensitivity checks without presenting regularisation as a substitute for information.

The Problem of Sparse Data in Regression

Classical maximum likelihood estimation can become unstable when sample sizes are small, events are rare, or predictors nearly separate outcome groups. In logistic regression, separation occurs when a predictor or predictor combination nearly or perfectly predicts the binary outcome. The fitted probabilities then approach 0 or 1, coefficient estimates become very large, and Wald standard errors stop being useful. A practical diagnostic is to inspect simple predictor-by-outcome tables for zero cells and to check whether the logistic model warns that fitted probabilities are numerically 0 or 1.

Penalised regression and Bayesian regression respond to the same problem in different language. Penalised regression adds a constraint to the likelihood so that extreme coefficients are pulled back toward more stable values. Bayesian regression combines the likelihood with prior distributions. Weakly informative priors serve as regularisation when the data alone are too thin to support precise estimates. Neither approach creates information that is not in the data. Both make the modelling assumptions more explicit.

Choosing Among Regularisation Strategies

Regularisation covers a family of strategies, and the appropriate choice depends on the outcome, the modelling goal, and the specific source of instability.

Problem in the small dataset	Better starting point	Why
Sparse binary outcome or separation warnings in logistic regression	Firth logistic regression	Produces finite bias-reduced estimates when ordinary maximum likelihood breaks down
Continuous outcome with correlated predictors	Ridge regression	Shrinks all slopes and reduces instability from collinearity without selecting a single "winner"
Continuous outcome with many candidate predictors and a screening goal	LASSO	Can set weak or unstable coefficients to zero, but selection should be treated as exploratory
Strong prior knowledge or a need to express assumptions directly	Bayesian regression with weakly informative priors	Regularises estimates through explicit prior distributions and requires convergence checks
Main goal is explanation of one pre-specified effect	Simpler pre-specified model plus sensitivity analysis	Penalised selection can obscure the estimand if the target effect was already known
Main goal is prediction	Penalised model with transparent tuning and validation	Prediction requires checking out-of-sample behaviour, not only coefficient significance

For small samples, the safest workflow is to fit the simplest scientifically meaningful model first, then use regularised estimates as sensitivity checks or as explicitly labelled prediction tools. If regularisation changes the substantive conclusion, report that instability rather than hiding it behind a single preferred model.

Firth-Penalised Logistic Regression

Firth's method reduces small-sample bias in likelihood estimation and is especially useful for sparse binary outcomes or separation (Firth 1993; Heinze and Schemper 2002). Table 13.1 shows the structure of a small project-success example. Most low-planning projects fail and most high-planning projects succeed, which creates a separation risk.

Table 13.1

Sparse project-success data used for the logistic-regression example

Planning band	No success	Success	Success rate
Planning score 1-5	11	1	8%
Planning score 6-9	0	8	100%

Note. The outcome is nearly separated by planning score. This is the setting where ordinary logistic regression can produce unstable estimates.

The ordinary logistic model fits in R, but the fitted probabilities are close to 0 or 1. That is a warning sign even if the software returns coefficients. Table 13.2 compares ordinary maximum-likelihood estimates with Firth-penalised estimates when the `logistf` package is available.

Table 13.2

Standard and Firth-penalised logistic-regression estimates

Method	Term	Estimate	Std. error	p-value
Standard ML	Planning score	43.85	42168.42	0.999
Standard ML	Prior experience	44.72	57464.52	0.999
Firth	Planning score	1.02	0.80	0.218
Firth	Prior experience	1.66	1.54	0.328

Note. The standard-model standard errors are Wald estimates from glm(); under separation they can become extremely large or effectively unbounded, especially when R warns that fitted probabilities are numerically 0 or 1. In this example, fitted probabilities close to 0 or 1 occurred: yes. Use logistf or another penalised method immediately when that warning appears. Firth's method adds a penalty proportional to the log determinant of the Fisher information matrix, reducing small-sample bias and preventing infinite estimates under separation. Coefficients remain on the log-odds scale and can be exponentiated for odds-ratio interpretation.

Interpretation should focus on direction, uncertainty, and model fragility. A finite Firth coefficient is a more stable estimate under a penalised likelihood, not confirmation that the effect size is precisely known. In small samples, report the event counts, variables in the model, penalisation method, confidence intervals, and whether ordinary logistic regression showed separation warnings.

Ridge Regression as Shrinkage

Ridge regression shrinks regression coefficients toward zero by adding a penalty proportional to the squared coefficient size. It is useful when predictors are correlated, sample size is modest, or the goal is prediction rather than unpenalised coefficient interpretation (Harrell 2015). Table 13.3 and Figure 13.1 show the same small customer-satisfaction regression under increasing ridge penalties.

Table 13.3

Ridge coefficient estimates under increasing penalty strength

Lambda	Intercept	Wait time	Friendliness
0	5.94	-0.24	1.14
1	5.94	-0.42	0.91
5	5.94	-0.51	0.68
20	5.94	-0.40	0.45

Note. Predictors were standardised to mean = 0 and SD = 1 before fitting because the L2 penalty is scale-dependent. Lambda = 0 is the ordinary least-squares solution. Coefficients apply to standardised predictors unless back-transformed.

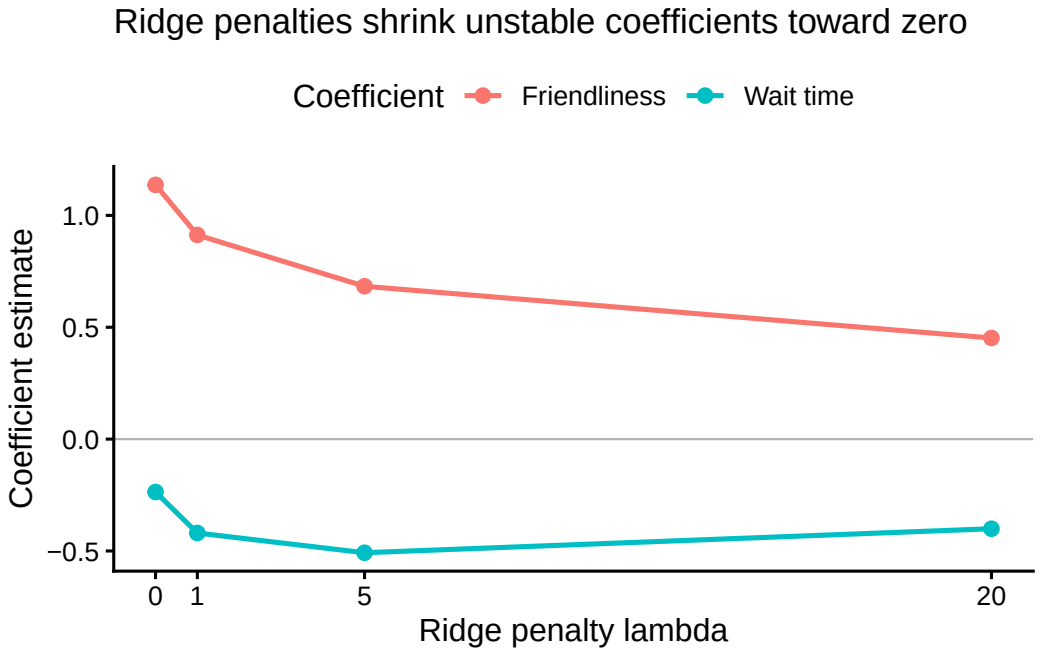


Figure 13.1: Ridge coefficient paths for the customer-satisfaction example.

As lambda increases, the slope estimates move toward zero. That shrinkage can reduce overfitting and improve prediction, but it also changes the estimand: the coefficients are penalised estimates, not ordinary least-squares coefficients. Report the penalty-selection method, whether predictors were standardised, and whether the model was used for prediction or interpretation.

Choosing the Ridge Penalty with `glmnet`

The code below shows the same ridge idea using `glmnet`, which is the package readers are most likely to use in practice. The predictors are standardised internally, `alpha = 0` requests ridge rather than lasso, and cross-validation selects a penalty. The sample is deliberately small, so the cross-validation curve should be read as a tuning aid rather than as a precise estimate of out-of-sample performance.

Table 13.4

Ridge estimates from glmnet at the selected penalty

Term	Estimate
(Intercept)	2.183
wait_time	-0.180
friendliness	0.740

Note. The selected lambda is the value minimising cross-validated prediction error. In small samples, repeat the analysis under plausible modelling choices rather than treating one cross-validation split as definitive.

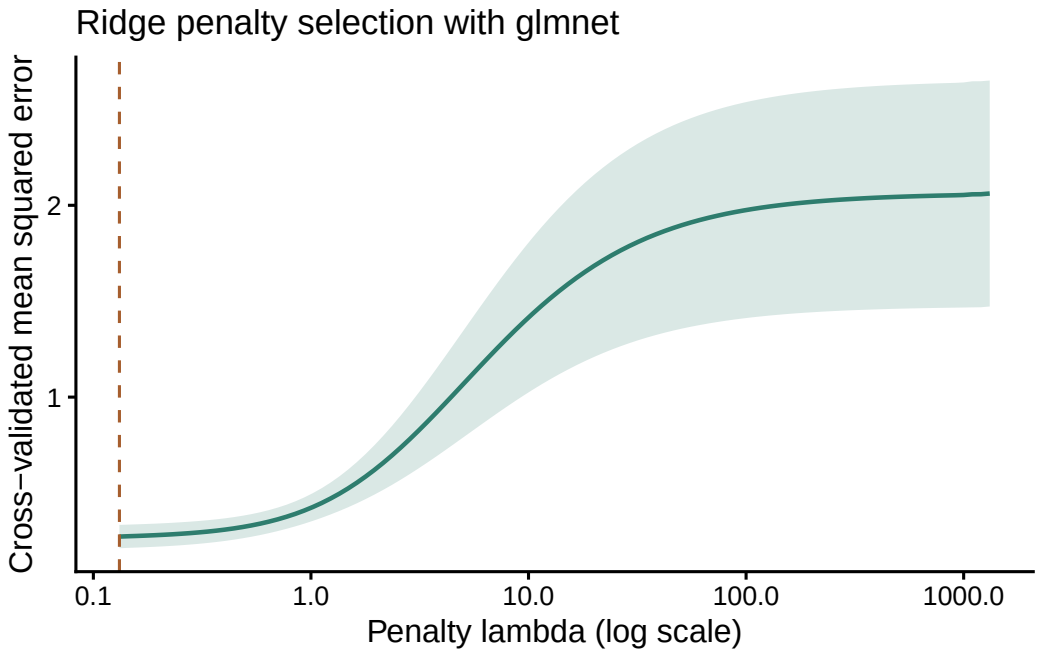


Figure 13.2: Cross-validation curve for ridge penalty selection with glmnet.

LASSO for Predictor Screening

The LASSO uses an L1 penalty rather than the squared L2 penalty used by ridge regression. This means that, as the penalty increases, some coefficients can be shrunk exactly to zero. That property makes LASSO useful for cautious predictor screening when a small dataset contains more candidate predictors than the sample can estimate reliably. It should not be treated as proof that excluded variables are irrelevant. With small samples, selected predictors can change under modest resampling or under a different set of candidate variables.

The example below uses 30 observations and five candidate predictors. Cross-validation chooses two common penalty values: `lambda.min`, which minimises the cross-validated error, and `lambda.1se`, which chooses a more parsimonious model within one standard error of that minimum.

Table 13.5

LASSO coefficients under two cross-validated penalty choices

Term	lambda.min	lambda.1se
Intercept	60.21	60.30
Service quality	3.10	2.27
Response speed	-1.61	-1.08
Price fairness	-0.81	0.00
Staff training	1.14	0.06
Waiting-room comfort	-0.19	0.00

Note. Coefficients were estimated with standardised predictors. Values equal to 0 indicate variables removed by the LASSO penalty at that tuning value.

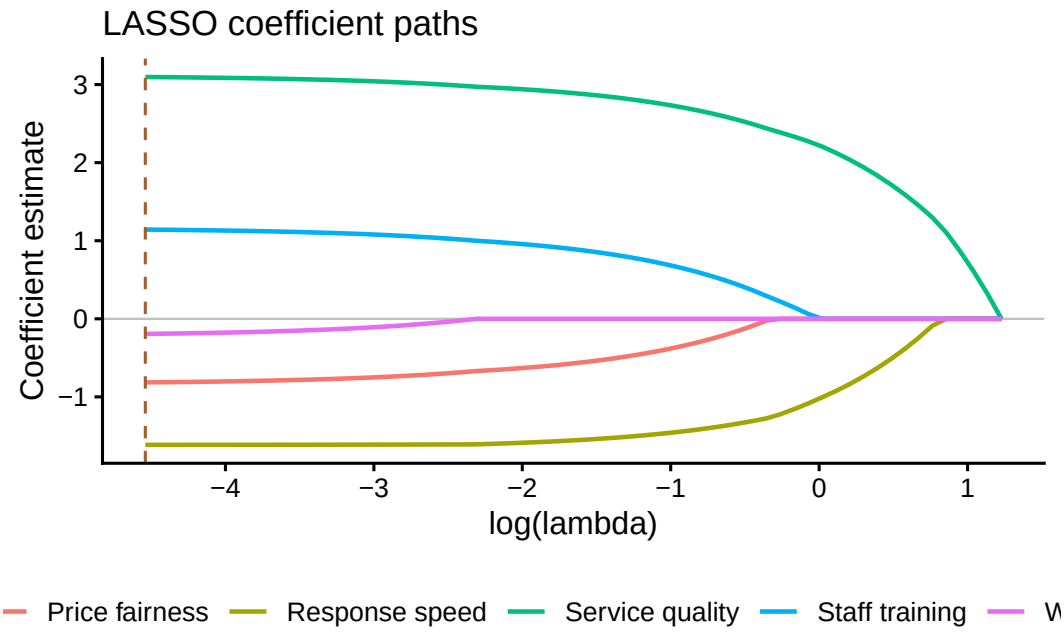


Figure 13.3: LASSO coefficient paths for five candidate predictors.

The main reporting point is the penalty rule, not just the final coefficients. If `lambda.1se` removes a predictor that `lambda.min` retains, describe that predictor as unstable rather than definitively absent. For explanatory work, LASSO is best used as a sensitivity analysis or screening tool before a simpler, pre-specified model is reported.

Bayesian Priors as Regularisation

Weakly informative priors are regularisation tools, not a way to force a preferred conclusion. A prior such as Normal(0, 2.5) on a logistic-regression coefficient implies that odds ratios between exp(-5) = 0.007 and exp(5) = 148 are plausible a priori while still regularising extreme estimates. Stronger priors such as Normal(0, 0.5) exert more shrinkage and should be justified by substantive knowledge or prior evidence (Gelman, Simpson, and Betancourt 2017).

Table 13.6 illustrates prior sensitivity for the wait-time slope in the customer-satisfaction example using a normal approximation to the likelihood. The table is not a replacement for full Bayesian computation, but it shows the main principle: when the prior is tight, the posterior moves toward zero. When the prior is weak, the posterior resembles the data-driven estimate.

Table 13.6

Prior-sensitivity illustration for a Bayesian regression slope

Prior on wait-time slope	Posterior mean	Posterior SD	95% credible interval
Normal(0, 0.5)	-0.17	0.27	-0.69 to 0.36
Normal(0, 1)	-0.21	0.30	-0.81 to 0.38
Normal(0, 2.5)	-0.23	0.32	-0.85 to 0.39
Normal(0, 10)	-0.24	0.32	-0.86 to 0.39

Note. The approximation uses the ordinary least-squares wait-time slope and standard error as a normal likelihood. Full Bayesian analyses should still check convergence diagnostics, such as R-hat < 1.01 and effective sample size > 400, and should inspect posterior predictive fit.

For a full Bayesian fit, use an MCMC package and report diagnostics. The following code is intentionally not evaluated in the book render because `brms` requires a working Stan toolchain, but it is the minimal workflow expected in a manuscript: specify priors, sample, check R-hat and effective sample size, and inspect posterior predictive fit. Before running it locally, verify the Stan toolchain with `cmdstanr::check_cmdstan_toolchain()` if using CmdStan, or confirm the installed `rstan` version with `rstan::stan_version()` before fitting a small test model.

```
library(brms)

small_prior <- c(
  prior(normal(0, 1), class = "b"),
  prior(student_t(3, 0, 2.5), class = "Intercept"),
  prior(exponential(1), class = "sigma")
)

bayes_fit <- brm(
  satisfaction ~ scale(wait_time) + scale(friendliness),
  data = customer_data,
  family = gaussian(),
  prior = small_prior,
  chains = 4,
  iter = 2000,
  seed = 2025
)

summary(bayes_fit)      # Check R-hat < 1.01 and bulk/tail ESS > 400
pp_check(bayes_fit)     # Inspect posterior predictive fit
```

Bayesian regression reports posterior intervals rather than frequentist confidence intervals. A 95% credible interval describes the range containing 95% of posterior probability given the model, data, and priors. That statement is conditional on the prior choice, so small-sample Bayesian reports should include the priors, convergence diagnostics, posterior predictive checks, and at least one plausible prior-sensitivity analysis. For leave-one-out cross-validation or WAIC

in Bayesian models, use them as approximate predictive checks, not as automatic proof that one small-sample model is correct (Vehtari, Gelman, and Gabry 2017).

Reporting Regularised Models

A regularised model report should make the stabilising assumption visible. State the outcome, sample size, event count where relevant, candidate predictors, standardisation, penalty or prior, tuning method and sensitivity checks. Do not report a penalised coefficient as if it were an ordinary unpenalised estimate.

For Firth logistic regression, report the sparse event table, the separation warning or diagnostic that motivated the method, coefficient scale, odds ratios if used, confidence intervals and software. For ridge and LASSO, report whether predictors were standardised, the value of lambda, how lambda was chosen, and whether conclusions change under `lambda.min` versus `lambda.1se` or under a simpler unpenalised model. For Bayesian models, report priors, seed, chains, iterations, R-hat, effective sample size, posterior intervals and posterior predictive checks.

A concise reporting sentence could read: "Because the ordinary logistic model produced fitted probabilities close to 0 and 1, we estimated a Firth-penalised logistic regression. The model included planning score and prior experience, reported coefficients on the log-odds scale with confidence intervals, and was interpreted as a stabilised sensitivity analysis rather than as precise evidence from a large sample." That level of detail is enough for readers to see both the method and the limitation.

Key Takeaways

Penalised and Bayesian regression methods are valuable in small samples because they make unstable estimates finite and reduce overfitting. Firth logistic regression is especially useful for sparse binary outcomes and separation. Ridge regression shrinks correlated or noisy linear-model coefficients. LASSO can screen candidate predictors by setting unstable coefficients to zero. Bayesian priors regularise estimates and express assumptions directly. The reporting obligation is the same in all cases: show the data structure, state the penalty or prior, check sensitivity, and avoid treating regularised estimates as more precise than the sample supports.

Self-Assessment Quiz

Question 1

What does separation mean in logistic regression?

a) Predictors are measured on different scales
b) A predictor or predictor combination nearly perfectly classifies the binary outcome

c) The residuals are normally distributed
d) The model has no intercept

Question 2

Why is Firth logistic regression useful with sparse binary outcomes?

a) It removes the binary outcome
b) It reduces small-sample bias and produces finite estimates under separation
c) It guarantees statistical significance
d) It replaces the need to report event counts

Question 3

What does ridge regression do to coefficients?

a) It forces all coefficients to be exactly zero
b) It shrinks coefficients toward zero by penalising large values
c) It converts a binary outcome into a count
d) It removes the need to standardise predictors

Question 4

In a small-sample Bayesian regression, why should prior sensitivity be reported?

a) Because priors have no effect
b) Because small samples can leave the posterior sensitive to reasonable prior choices
c) Because it replaces convergence diagnostics
d) Because credible intervals are always narrow

Question 5

Which statement is the safest interpretation of a regularised estimate?

a) It is automatically unbiased
b) It is stabilised by an explicit penalty or prior and should be interpreted with that assumption stated
c) It proves that the effect exists
d) It needs no confidence or credible interval

Answers and Explanations

Question 1

Answer: b)

Explanation: Separation occurs when the outcome can be nearly or perfectly predicted from the covariates. Standard logistic maximum likelihood can then produce extremely large or infinite coefficients.

Question 2

Answer: b)

Explanation: Firth's method uses a penalised likelihood that reduces small-sample bias and avoids infinite estimates. Event counts and uncertainty still need to be reported.

Question 3

Answer: b)

Explanation: Ridge regression adds a squared-coefficient penalty. Larger penalties pull estimates toward zero and can reduce overfitting.

Question 4

Answer: b)

Explanation: When the data are thin, different plausible priors can lead to noticeably different posterior estimates. Reporting sensitivity shows how much the conclusion depends on modelling assumptions.

Question 5

Answer: b)

Explanation: Regularisation stabilises estimates by adding assumptions. Those assumptions, along with intervals and diagnostics, must be reported.

Chapter 14: Multi-Criteria Decision Making (MCDM) for Small Sets of Alternatives

Learning Objectives

By the end of this chapter, you will be able to distinguish decision analysis from statistical inference, structure a small multi-criteria decision problem, compute transparent weighted rankings with AHP and TOPSIS, check consistency and sensitivity, and report rankings without implying more precision than stakeholder judgements support.

When MCDM Methods Are Appropriate

Multi-criteria decision-making methods are useful when a small set of alternatives must be ranked or selected using several criteria. Their value lies in transparency: the criteria, weights, normalisation choices, and aggregation rule are all visible, so stakeholders can see how a ranking was produced and whether it holds under plausible assumptions. They produce ranked or weighted scores rather than p-values or population-parameter estimates.

MCDM belongs in a small-sample methods text because many constrained research settings end with a decision rather than a population estimate. A school may need to choose one of three reading interventions, a clinic may need to prioritise one of four service improvements, or a community project may need to allocate a small grant among a few feasible options. When the alternatives are fixed and the evidence is too limited for strong inference, a structured decision model is more honest than pretending that a p-value can select the best option.

MCDM is appropriate when alternatives are few, criteria are heterogeneous, stakeholder preferences matter, and the goal is selection or resource allocation (Saaty 1980; Hwang and Yoon 1981). When the question is whether a treatment caused an effect, only a controlled experiment can answer it.

Analytic Hierarchy Process

The Analytic Hierarchy Process (AHP) uses pairwise comparisons to derive priority weights (Saaty 1980). Decision-makers compare criteria two at a time, and the resulting matrix is converted into weights. AHP also includes a consistency check. A low consistency ratio suggests

that the pairwise judgements are coherent enough to use. A high value means the comparisons should be revisited.

Before constructing the matrix, document the elicitation protocol: who supplied the comparisons, what scale was used, whether judgements were individual or consensus-based, and how disagreements were resolved. A simple stakeholder template should ask each rater to compare every pair of criteria, give a short reason for each judgement, and flag any comparison they feel uncertain about. The final matrix should be auditable rather than treated as a hidden expert input.

Table 14.1 gives the criteria weights for a training-programme selection example. Effectiveness receives the largest weight, while cost and feasibility still contribute to the decision.

The AHP calculation below uses the principal eigenvector of the pairwise-comparison matrix. The consistency ratio is computed as `(lambda_max - k) / ((k - 1) * RI)`, where `k` is the number of criteria and `RI` is Saaty's random-index value for a matrix of that size.

```
criteria_names <- c("Cost", "Effectiveness", "Feasibility")
ahp_matrix <- matrix(
  c(
    1,   1/3, 1/1.5,
    3,   1,   2,
    1.5, 1/2, 1
  ),
  nrow = 3,
  byrow = TRUE,
  dimnames = list(criteria_names, criteria_names)
)

eig <- eigen(ahp_matrix)
weights <- Re(eig$vectors[, 1])
weights <- weights / sum(weights)

lambda_max <- Re(eig$values[1])
consistency_index <- (lambda_max - length(criteria_names)) /
 ↪  (length(criteria_names) - 1)
consistency_ratio <- consistency_index / 0.58
```

Table 14.1

AHP criteria weights for the training-programme decision

Criterion	Weight	Relative emphasis
Cost	0.182	Cost matters, but is not dominant
Effectiveness	0.545	Primary criterion
Feasibility	0.273	Secondary practical criterion

Note. The consistency ratio is 0.000. A CR below 0.10 is commonly treated as acceptable for pairwise-comparison matrices (Saaty, 1980). For very small 3 x 3 matrices, some authors tolerate slightly higher values, but inconsistent judgements should still be revisited.

Table 14.2

AHP weighted scores for three training programmes

Rank	Programme	Weighted cost	Weighted effectiveness	Weighted feasibility	Total score
1	B	0.055	0.273	0.082	0.409
2	A	0.091	0.136	0.109	0.336
3	C	0.036	0.136	0.082	0.255

Note. Programme B ranks highest because effectiveness has the largest weight and Programme B performs best on that criterion.

Interpretation should be decision-focused rather than inferential. Programme B is the preferred option under the stated weights and scores. That does not mean Programme B is statistically superior. It means the explicit decision model ranks it highest.

TOPSIS

TOPSIS ranks alternatives by closeness to an ideal solution and distance from a negative-ideal solution (Hwang and Yoon 1981). The method first normalises the criteria, applies weights, identifies the best and worst value on each weighted criterion, and then computes a closeness coefficient. Higher coefficients indicate alternatives closer to the ideal profile.

Cost criteria require recoding because TOPSIS assumes larger values are better. In Table 14.3, project cost is transformed into a benefit score before normalisation.

The vector-normalisation step divides each criterion value by the Euclidean norm of its column:

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_i x_{ij}^2}}.$$

The implementation below shows the full calculation, including the cost-to-benefit transformation and closeness coefficient.

```
decision_matrix <- tibble(
  Project = c("P1", "P2", "P3", "P4"),
  Impact = c(7, 8, 6, 9),
  Cost = c(50, 70, 40, 80),
  Support = c(6, 7, 8, 7)
)

weights_topsis <- c(Impact = 1/3, Cost_benefit = 1/3, Support = 1/3)

topsis_work <- decision_matrix |>
  mutate(Cost_benefit = max(Cost) - Cost + min(Cost))

normalised <- topsis_work |>
  mutate(across(c(Impact, Cost_benefit, Support),
                ~ .x / sqrt(sum(.x^2)),
                .names = "norm_{.col}"))

weighted <- normalised |>
  mutate(
    w_Impact = norm_Impact * weights_topsis["Impact"],
    w_Cost = norm_Cost_benefit * weights_topsis["Cost_benefit"],
    w_Support = norm_Support * weights_topsis["Support"]
  )

ideal <- c(max(weighted$w_Impact), max(weighted$w_Cost),
 ↪  max(weighted$w_Support))
negative_ideal <- c(min(weighted$w_Impact), min(weighted$w_Cost),
 ↪  min(weighted$w_Support))

topsis_results <- weighted |>
  rowwise() |>
  mutate(
    distance_ideal = sqrt(sum((c(w_Impact, w_Cost, w_Support) - ideal)^2)),
    distance_negative = sqrt(sum((c(w_Impact, w_Cost, w_Support) -
     ↪  negative_ideal)^2)),
    closeness = distance_negative / (distance_ideal + distance_negative)
  ) |>
  ungroup() |>
  arrange(desc(closeness))
```

Table 14.3

TOPSIS ranking for four community projects

Rank	Project	Impact	Cost	Support	Closeness coefficient
1	P3	6	40	8	0.640

Rank	Project	Impact	Cost	Support	Closeness coefficient
2	P1	7	50	6	0.544
3	P2	8	70	7	0.395
4	P4	9	80	7	0.389

Note. Cost was recoded as a benefit before vector normalisation, where each criterion value is divided by the Euclidean norm of its column. This places criteria on a comparable scale before weighting. The closeness coefficient ranges from 0 to 1.

The TOPSIS ranking depends on the normalisation rule and the weights. If stakeholders disagree about cost or impact weights, the ranking should be recomputed under those alternatives rather than reported as a single unquestioned answer.

VIKOR and Other MCDM Methods

VIKOR is another ideal-solution method, but it emphasises compromise between group utility and individual regret (Opricovic and Tzeng 2004). TOPSIS asks which alternative is geometrically closest to the ideal profile. VIKOR asks which alternative is a defensible compromise when no option is best on every criterion. Other methods, such as SMART, WASPAS, and MOORA, follow the same broad logic: structure the decision, normalise criteria, apply weights, aggregate scores, and test sensitivity.

The VIKOR calculation below uses the same four projects and equal criterion weights as the TOPSIS example. For each criterion, the best value receives zero loss and worse values receive larger normalised loss. The utility measure `S` summarises total weighted loss, the regret measure `R` records the largest single-criterion loss, and the index `Q` combines both using `v = 0.5` to balance group utility and individual regret. Lower `Q` values are preferred.

```
vikor_scaled <- topsis_work |>
  transmute(
    Project,
    Impact = weights_topsis["Impact"] *
      (max(Impact) - Impact) / (max(Impact) - min(Impact)),
    Cost_benefit = weights_topsis["Cost_benefit"] *
      (max(Cost_benefit) - Cost_benefit) / (max(Cost_benefit) -
↪   min(Cost_benefit)),
    Support = weights_topsis["Support"] *
      (max(Support) - Support) / (max(Support) - min(Support))
  ) |>
  mutate(
    S = Impact + Cost_benefit + Support,
    R = pmax(Impact, Cost_benefit, Support),
    Q = 0.5 * (S - min(S)) / (max(S) - min(S)) +
      0.5 * (R - min(R)) / (max(R) - min(R))
```

```
  ) |>
  arrange(Q)
```

Table 14.4

VIKOR compromise ranking for the community project example

Rank	Project	Utility loss (S)	Regret loss (R)	VIKOR index (Q)
1	P2	0.528	0.250	0.318
2	P3	0.333	0.333	0.500
3	P4	0.500	0.333	0.773
4	P1	0.639	0.333	1.000

Note. Lower Q values are preferred. Compare this compromise ranking with the TOPSIS closeness ranking. Disagreement is a sensitivity finding, not an error.

The technical differences matter less than the reporting discipline. Always state the criteria and their substantive justification, the raw scores and direction of preference, the weights and how they were elicited, the normalisation and aggregation methods, and sensitivity analyses across plausible weight sets. Weights are value judgements, not statistical estimates, so they need transparent justification.

Sensitivity Analysis

Sensitivity analysis is essential because MCDM rankings can change when weights or normalisation rules change. Figure 14.1 and Table 14.5 vary the weight placed on effectiveness in the AHP example. This is a better way to communicate robustness than presenting a ranking as if it were fixed by the data alone.

The sensitivity loop below perturbs the effectiveness weight and recomputes rankings. In a stakeholder report, repeat this for each contested criterion or use a tornado plot to show rank reversals under +/-20% weight changes.

```
sensitivity_results <- purrr::map_dfr(seq(0.30, 0.80, by = 0.05),
 ↪  function(eff_weight) {
  remaining <- 1 - eff_weight
  cost_weight <- remaining * (criteria_weights["Cost"] /
    (criteria_weights["Cost"] + criteria_weights["Feasibility"]))
  feas_weight <- remaining * (criteria_weights["Feasibility"] /
    (criteria_weights["Cost"] + criteria_weights["Feasibility"]))

  scores <- programme_scores |>
    transmute(
      Programme,
```

```
      Score = Cost * cost_weight +
        Effectiveness * eff_weight +
        Feasibility * feas_weight
    ) |>
    arrange(desc(Score))

  tibble(
    effectiveness_weight = eff_weight,
    top_programme = scores$Programme[1],
    ranking = paste(scores$Programme, collapse = " > ")
  )
})
```

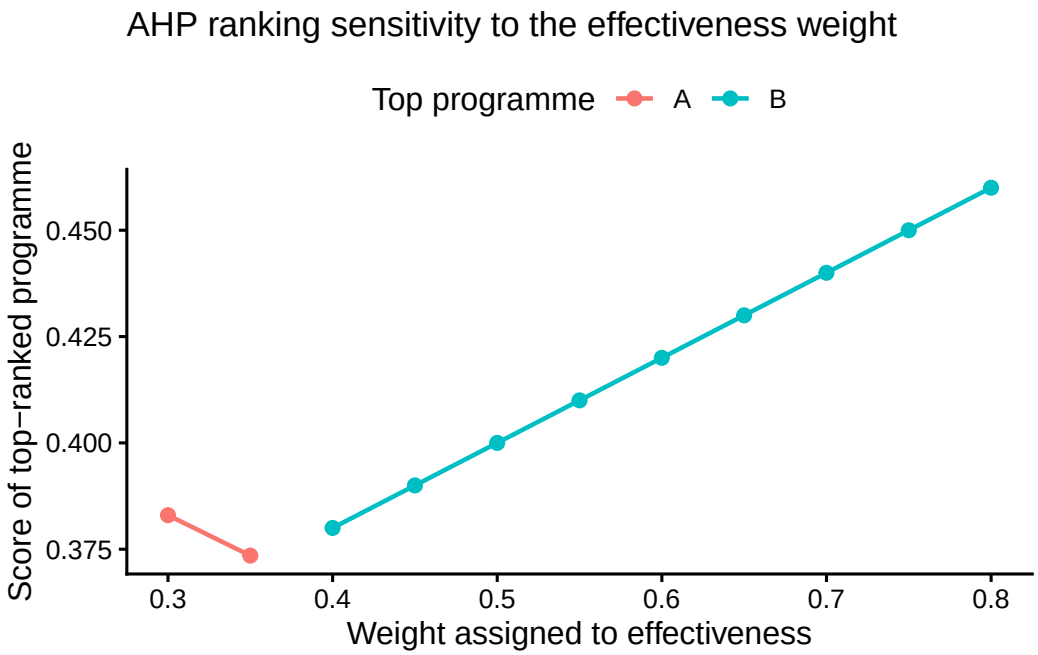


Figure 14.1: Sensitivity of the AHP top-ranked programme to the effectiveness weight.

For applied decision reports, a tornado plot can extend this idea by varying each criterion weight by a fixed amount, such as +/-20%, and showing whether the top-ranked alternative changes. That display is often more comprehensive than a single-weight sweep when several criteria are contested.

Table 14.5

AHP ranking sensitivity as the effectiveness weight changes

Effectiveness weight	Top programme	Ranking
0.30	A	A > B > C
0.35	A	A > B > C
0.40	B	B > A > C
0.45	B	B > A > C
0.50	B	B > A > C
0.55	B	B > A > C
0.60	B	B > A > C
0.65	B	B > A > C
0.70	B	B > A > C
0.75	B	B > A > C
0.80	B	B > A > C

Note. The ranking is stable when Programme B remains first across plausible effectiveness weights. If the top programme changes, stakeholders should discuss whether the weight range is realistic.

Key Takeaways

MCDM methods help structure decisions with multiple criteria and few alternatives. AHP is useful when pairwise judgements are central. TOPSIS and VIKOR are useful when alternatives can be compared against ideal and compromise profiles. These methods complement statistical inference but do not replace it. Good MCDM reporting makes the judgement calls visible: criteria, scores, weights, normalisation, aggregation, consistency, and sensitivity all need to be shown.

Self-Assessment Quiz

Question 1

When are MCDM methods most appropriate?

a) When the goal is hypothesis testing
b) When a small set of alternatives must be ranked using multiple criteria
c) When there is only one outcome variable
d) When p-values are required

Question 2

What does AHP's consistency ratio evaluate?

a) Whether residuals are normally distributed
b) Whether pairwise judgements are coherent enough to use
c) Whether the sample size is large
d) Whether every criterion has equal weight

Question 3

In TOPSIS, what does a larger closeness coefficient mean?

a) The alternative is closer to the ideal and farther from the negative ideal
b) The p-value is smaller
c) The criterion is less important
d) The alternative has a larger sample size

Question 4

Why must cost criteria often be transformed before TOPSIS?

a) TOPSIS cannot use numbers below zero
b) TOPSIS assumes larger criterion values are better
c) Cost is always irrelevant
d) It is required for statistical significance

Question 5

Why is sensitivity analysis essential in MCDM?

a) It replaces stakeholder judgement
b) It shows whether rankings change under plausible weights or inputs
c) It computes exact p-values
d) It removes the need to state criteria

Answers and Explanations

Question 1

Answer: b)

Explanation: MCDM methods support ranking and selection decisions across multiple criteria. They are decision tools, not hypothesis tests.

Question 2

Answer: b)

Explanation: The consistency ratio checks whether pairwise comparisons are logically coherent. High inconsistency means the judgements should be revisited.

Question 3

Answer: a)

Explanation: TOPSIS ranks alternatives by relative closeness to the ideal solution. Larger coefficients indicate more preferred alternatives under the chosen weights and normalisation.

Question 4

Answer: b)

Explanation: TOPSIS treats larger values as preferable. Cost must therefore be converted so that lower cost becomes a higher benefit score.

Question 5

Answer: b)

Explanation: Rankings can be sensitive to weights and normalisation. Sensitivity analysis shows whether the preferred alternative is robust or depends on a narrow assumption.

Summary of Part C

Part C presented the main analytic toolkit for small-sample quantitative work. Across these chapters, the emphasis was on choosing methods that remain defensible when sample sizes are modest, events are rare, distributions are skewed, predictors are numerous, or the set of alternatives is small. Exact and resampling methods, rank-based procedures, sparse-count methods, short time-series approaches, penalised and Bayesian regression, and multi-criteria decision making all served the same purpose: matching the method to the information actually available.

Taken together, these chapters show that small samples do not call for one universal workaround. They call for a careful match between the research question, the structure of the data, the assumptions that can be defended, and the kind of uncertainty that needs to be reported.

Method-Selection Framework

Use this table as a starting point before choosing a procedure. It does not replace design knowledge, but it helps connect common small-sample questions to the chapters that give the worked implementation.

Data structure or question	Typical small-sample setting	Recommended starting point	Main chapter
Two categorical variables with sparse cells	2 x 2 table, expected counts below 5	Fisher's exact test, with mid-p or unconditional exact tests as sensitivity checks	Chapter 10
One sparse event count against a benchmark	Few events observed over a known exposure	Exact Poisson test and exact confidence interval	Chapter 10
Custom statistic or skewed continuous outcome	n is too small for stable normal approximation	Permutation test when exchangeability is defensible, or bootstrap interval when resampling the statistic is meaningful	Chapter 10

Data structure or question	Typical small-sample setting	Recommended starting point	Main chapter
Two independent ordinal or skewed groups	Ratings, waiting times, or outcomes with outliers	Mann–Whitney U with medians, IQRs, and Cliff's delta	Chapter 11
Paired ordinal or skewed measurements	Before-after or matched-pair design	Wilcoxon signed-rank with the Hodges–Lehmann pseudomedian shift	Chapter 11
Three or more independent groups	Small groups with ordinal or skewed outcomes	Kruskal–Wallis with adjusted pairwise follow-up and epsilon-squared	Chapter 11
Repeated conditions for the same participants	Three or more within-person conditions	Friedman test with Kendall's W and adjusted paired follow-up	Chapter 11
Sparse rates or count outcomes	Few events, overdispersion, or short exposure summaries	Exact rate methods first, then quasi-Poisson or negative binomial sensitivity checks when model assumptions are defensible	Chapter 12
Very short time series	12 to 20 ordered observations with a simple trend question	Transparent trend model with residual-bootstrap forecast band only if residuals are approximately independent	Chapter 12
Sparse binary regression or separation	Few events or fitted probabilities near 0 or 1	Firth logistic regression and clear event-count reporting	Chapter 13
Many or correlated predictors relative to n	Linear model with unstable slopes	Ridge regression for shrinkage, LASSO for screening, and standardisation before penalisation	Chapter 13

Data structure or question	Typical small-sample setting	Recommended starting point	Main chapter
Bayesian small-sample model	Prior information is explicit and defensible	Weakly informative priors, convergence diagnostics, posterior predictive checks, and prior sensitivity	Chapter 13
Ranking a few alternatives rather than testing a sample effect	Fixed set of programmes, suppliers, or designs	AHP, TOPSIS, or VIKOR with transparent weights and sensitivity analysis	Chapter 14

Part D: Reporting and Interpretation

This part addresses how to communicate findings from small-sample studies transparently and responsibly. We cover effect sizes and confidence intervals, transparent reporting of methods and limitations, interpreting non-significant results, presenting uncertainty visually, and documenting analytic choices.

In This Part

- Chapter 15: Effect Sizes and Confidence Intervals over P-Values
- Chapter 16: Interpreting Non-Significant Results
- Chapter 17: Transparent Reporting of Methods and Limitations
- Chapter 18: Visualising Uncertainty and Presenting Results
- Summary of Part D

Chapter 15: Effect Sizes and Confidence Intervals over P-Values

Learning Objectives

By the end of this chapter, you will be able to explain why p-values alone are insufficient for small-sample inference, compute and interpret common effect-size measures, use confidence intervals to judge magnitude and precision, and report results in language that separates statistical evidence from practical importance.

Why P-Values Are Not Enough

A p-value answers a narrow question: how unusual the data would be if the null hypothesis and modelling assumptions were true. It does not tell the reader whether an effect is large, precise, clinically important, educationally meaningful, or worth acting on. In small studies, this limitation is especially visible because even practically important effects can have p-values above 0.05 when the confidence interval is wide.

Effect sizes and confidence intervals restore the missing information. The effect size describes magnitude, while the interval describes the range of values that remain compatible with the data. A transparent report should therefore make the p-value secondary to the estimate, its uncertainty, and its practical interpretation.

Common Effect-Size Metrics

Cohen's d expresses a mean difference in pooled standard-deviation units. Odds ratios compare the odds of a binary outcome between groups. Correlations such as Pearson's r, Spearman's rho, and Kendall's tau describe association. Variance-accounted-for measures such as eta-squared and epsilon-squared are useful for omnibus comparisons. These metrics are not interchangeable; the correct choice depends on the outcome scale, design, and question.

For ANOVA-style summaries, eta-squared is often biased upward in small samples because it attributes sample-specific noise to the effect. Epsilon-squared and omega-squared are usually more cautious alternatives. For a Kruskal–Wallis test, a common epsilon-squared estimate is

$\epsilon^2 = (H - k + 1)/(n - k)$, where H is the test statistic, k is the number of groups, and n is the total sample size (Tomczak and Tomczak 2014).

Cohen's conventional benchmarks of d = 0.20, 0.50, and 0.80 are useful only as a starting vocabulary (Cohen 1988). Practical importance is domain-specific. In education, effects as small as d = 0.10 can be meaningful when scaled to large populations (Kraft 2020); in clinical trials, even d = 0.50 may be insufficient if the intervention carries substantial risk, cost, or burden.

Table 15.1

Context-specific effect-size benchmarks

Domain	Small benchmark	Moderate benchmark	Large benchmark	Source
Psychological interventions	d = 0.20	d = 0.50	d = 0.80	Cohen; Lipsey and Wilson
Educational interventions	d = 0.05-0.10	d = 0.20-0.30	d >= 0.50	Kraft
Medical treatments	OR = 1.20	OR = 2.00	OR >= 4.00	Chen and colleagues
Customer satisfaction	d = 0.20	d = 0.50	d = 0.80	Domain standard
Process improvements	5% change	10-20% change	>=25% change	Domain standard

Note. The values are interpretive anchors, not universal cut-offs. Use domain evidence and substantive consequences when judging magnitude.

Educational effects are often smaller than laboratory effects but can still matter when an intervention reaches many students (Kraft 2020). Medical odds ratios require clinical context because the same odds ratio can imply very different absolute risk changes depending on baseline risk (Chen, Cohen, and Chen 2010). Meta-analytic norms, such as those reviewed by Lipsey and Wilson (1993), are useful because they place a new estimate within an empirical distribution of related studies.

Mean Differences and Standardised Effects

The most interpretable effect size is often the unstandardised mean difference. In the teaching-method example below, Method A has a higher average score than Method B. The difference in original score units is the first result to report; Cohen's d and Hedges' g help compare the result with other studies that used different scales.

Table 15.2

Mean-difference and standardised effect-size summary

Quantity	Estimate	95% CI	Interpretation
Method A mean	81.5	–	Average score under Method A
Method B mean	79.8	–	Average score under Method B
Mean difference (A - B)	1.8	-1.0 to 4.5	Difference in original test-score units
Cohen's d	0.54	-0.32 to 1.40	Standardised difference using the pooled SD
Hedges' g	0.52	-0.31 to 1.35	Small-sample correction to Cohen's d

Note. Two independent groups with n = 12 per group. Equal-variance t test: t(22) = 1.32, p = 0.200.

Students in Method A scored about 1.8 points higher than those in Method B, with a 95% CI from -1.0 to 4.5 points. On the standardised scale, Cohen's d is about 0.54 and Hedges' g is about 0.52. The reporting lesson is to give the reader both the original-unit difference and the standardised comparison, while recognising that the standardised interval remains wide with only 12 observations per group.

Binary Effects: Odds Ratios, Risk Differences, and NNT

For binary outcomes, odds ratios are common but can be difficult to interpret. Odds ratios compare odds, not risks. When baseline risk is low, roughly below 10%, the odds ratio approximates the risk ratio; when baseline risk is higher, the odds ratio can substantially overstate the apparent effect. Risk differences are often more directly useful because they describe the absolute change in event probability. The number needed to treat (NNT) is the reciprocal of the absolute risk difference and answers how many people would need the intervention for one additional success relative to control.

Table 15.3

Binary effect-size summaries for a small trial

Quantity	Estimate	95% CI	Interpretation
Treatment success rate	0.75	–	15 successes among 20 treated participants
Control success rate	0.50	–	10 successes among 20 control participants
Absolute risk difference	0.25	-0.04 to 0.54	Additional success probability under treatment
Odds ratio	2.92	0.66 to 14.50	Odds of success in treatment relative to control

Quantity	Estimate	95% CI	Interpretation
Number needed to treat	4.0	Not bounded because the risk-difference CI crosses 0	Approximate patients treated for one additional success

Note. Fisher's exact test for the 2 x 2 table gives two-sided p = 0.191. The odds-ratio CI is exact; the odds ratio is the conditional maximum-likelihood estimate from Fisher's exact test and can differ slightly from the simple cross-product ratio. The risk-difference CI is an uncorrected large-sample interval used here for interpretation.

The treatment group has a higher observed success rate, but the confidence interval for the risk difference crosses zero and the exact odds-ratio interval is wide. Because NNT is the reciprocal of the risk difference, an interval that includes zero crosses infinity and produces an unbounded or discontinuous NNT interval. In such cases, report the risk difference and its interval as the primary inferential summary, and present the point-estimate NNT with a cautionary note rather than as a simple bounded interval.

Confidence Intervals as the Primary Summary

Confidence intervals are often more informative than p-values because they communicate magnitude and precision simultaneously. An interval that excludes the null value also corresponds to a statistically significant result at the matching alpha level, but its real value is broader: it shows whether the data are compatible with effects that would matter in practice.

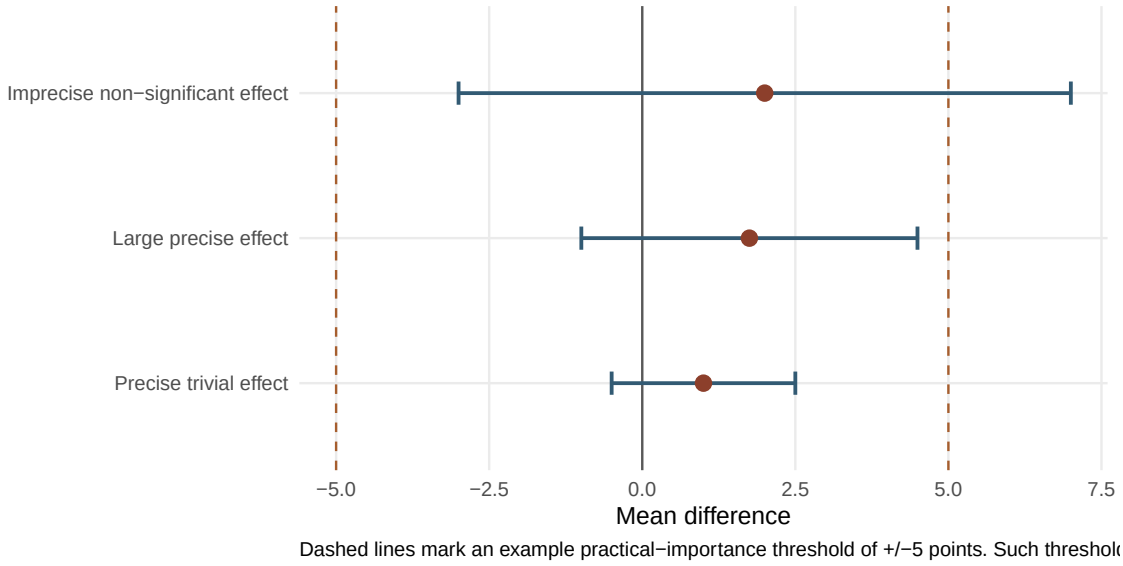


Figure 15.1: Confidence intervals communicate both magnitude and precision.

In small-sample work, a non-significant result with a wide interval should usually be described as imprecise rather than negative. Conversely, a narrow interval that rules out practically important effects can support a more substantive conclusion even when p is above 0.05. Chapter 16 develops this distinction in detail.

Reporting Effect Sizes

A complete results sentence names the estimator, gives the point estimate, reports the confidence interval, and then interprets the magnitude in context. For the teaching-method example, a concise report would be: "Method A scores were higher than Method B scores by 1.8 points, 95% CI [-1.0, 4.5], corresponding to Hedges' g = 0.52." The practical interpretation should then explain whether a 1.8-point gain is educationally meaningful, not merely whether the p-value crossed a threshold.

Results Sentence Builder

Use this structure when writing small-sample results:
Estimator: Name the comparison or association in original units.
Magnitude: Report the point estimate and a suitable effect size.
Uncertainty: Report the confidence interval before the p-value if both are included.
Context: State the practical threshold, benchmark, or prior evidence used to judge importance.
Caution: If the interval is wide, say that the estimate is imprecise rather than overinterpreting the point estimate.
Example: "The intervention group scored 4.2 points higher than the comparison group, 95% CI [0.8, 7.6], Hedges' g = 0.62. The interval includes values from a small educational benefit to a larger practically meaningful gain, so the estimate should be treated as promising but still imprecise."

When adapting the code examples, record the random seed for any simulation or bootstrap step and include package versions in reproducible reports. Stochastic procedures can otherwise produce slightly different intervals across machines.

Key Takeaways

P-values alone are too thin for small-sample interpretation. Effect sizes describe how large a result is, confidence intervals describe how precisely it has been estimated, and original-unit summaries often communicate practical importance better than standardised metrics. A strong report gives the estimate, its uncertainty, and the substantive benchmark used to judge whether the effect matters.

Self-Assessment Quiz

Question 1

Cohen's d = 0.50 means:

a) The p-value is 0.50
b) The means differ by one half of a pooled standard deviation
c) Half of participants improved
d) The effect explains 50% of the variance

Question 2

Why should effect sizes be reported alongside p-values?

a) Effect sizes quantify magnitude, whereas p-values mainly summarise evidence against a null value
b) Effect sizes make confidence intervals unnecessary
c) Effect sizes are useful only when $p < 0.05$
d) Effect sizes remove sampling uncertainty

Question 3

An odds ratio of 3.0 means:

a) The outcome is exactly three times as common
b) The odds of the outcome are three times as high in one group
c) The risk difference is 3 percentage points
d) The p-value must be below 0.05

Question 4

A confidence interval for a mean difference is -2 to 8 points. What is the best interpretation?

a) The null hypothesis is true
b) The result is inconclusive because plausible values range from slightly negative to meaningfully positive
c) The effect is definitely harmful
d) The confidence interval is irrelevant if $p > 0.05$

Question 5

Why can an NNT confidence interval be difficult to report when the risk-difference interval crosses zero?

a) The NNT is undefined at zero risk difference
b) Fisher's exact test cannot be used
c) The odds ratio must equal 1
d) NNT is only for continuous outcomes

Question 6

Which statement about generic benchmarks such as d = 0.20, 0.50, and 0.80 is best?

a) They are universal cut-offs
b) They are conventions that must be interpreted in context
c) They prove clinical importance
d) They apply only to large samples

Answers and Explanations

Question 1

Answer: b)

Explanation: Cohen's d is a standardised mean difference. A value of 0.50 means the two group means differ by one half of the pooled standard deviation.

Question 2

Answer: a)

Explanation: A p-value does not tell the reader how large an effect is. The effect size supplies the magnitude, and the confidence interval supplies the uncertainty around that magnitude.

Question 3

Answer: b)

Explanation: Odds ratios compare odds, not risks. When outcomes are common, an odds ratio can differ meaningfully from the corresponding risk ratio.

Question 4

Answer: b)

Explanation: The interval includes zero but also includes positive values that may be practically important. That pattern indicates imprecision, not proof of no effect.

Question 5

Answer: a)

Explanation: NNT is the reciprocal of the risk difference. If the risk-difference interval includes zero, the reciprocal crosses infinity, so a simple bounded NNT interval is misleading.

Question 6

Answer: b)

Explanation: Generic benchmarks are useful shorthand, but practical importance depends on the outcome, cost, risk, and prior evidence in the field.

Chapter 16: Interpreting Non-Significant Results

Learning Objectives

By the end of this chapter, you will be able to distinguish “no evidence of effect” from “evidence of no meaningful effect,” interpret non-significant results using confidence intervals and power, recognise when a study is inconclusive, and use equivalence or non-inferiority logic when the research question is about similarity rather than difference.

What a Non-Significant Result Means

A non-significant result means that the observed data were not sufficiently inconsistent with the null hypothesis to reject it at the chosen alpha level. It does not prove the null hypothesis, and it does not show that an intervention is ineffective. With small samples, non-significant results are common because power is limited; a study with 30% power will fail to reject the null 70% of the time even when the alternative hypothesis is true.

The phrase “absence of evidence is not evidence of absence” is useful here, but it should be qualified by power. If a design has only 30% power to detect a meaningful effect, a non-significant result provides little information either way. If a study reports $p = 0.15$, the safest conclusion is that the data do not provide strong evidence against the null. Whether the result also rules out practically meaningful effects depends on the confidence interval.

Reading the Confidence Interval

The confidence interval is the main tool for interpreting a non-significant result. A narrow interval that excludes effects large enough to matter can support a claim that any remaining effect is likely trivial. A wide interval that includes trivial and important values means the study is inconclusive.

Table 16.1

Confidence-interval interpretations for non-significant results

Scenario	n per group	Mean difference	95% CI	p-value	Interpretation
Narrow non-significant interval	100	1	-0.5 to 2.5	0.120	If differences below 3 points are trivial, this interval rules out a meaningful benefit.
Wide non-significant interval	12	2	-3.0 to 7.0	0.350	The interval includes harmful, trivial, and beneficial values; the study is inconclusive.
Equivalence not established	12	2	-3.0 to 7.0	0.450	The interval extends outside the +/-5 equivalence margin, so equivalence is not demonstrated.

Note. All examples use a mean-difference scale where positive values favour the intervention.

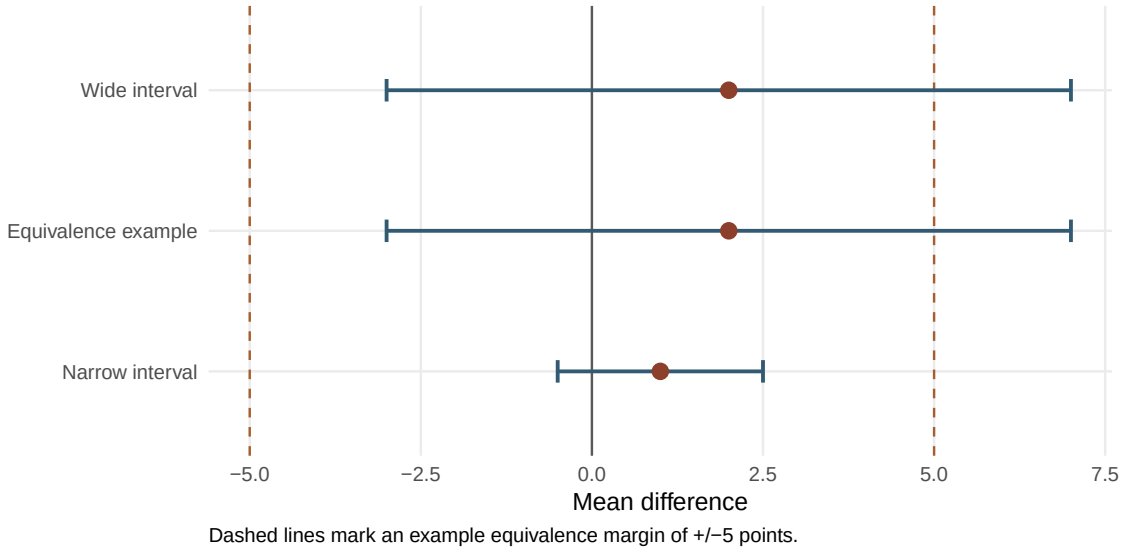


Figure 16.1: Non-significant results can be precise, inconclusive, or insufficient for equivalence.

The first interval rules out effects above a practical threshold of 3 points, so a cautious "no meaningful difference" interpretation may be reasonable if that threshold was defined in advance. The second interval is wider and includes values that could change the substantive conclusion. The third example shows why a small point estimate alone is not enough to claim equivalence.

Power and Minimum Detectable Effects

Small samples can fail to detect effects that would matter in practice. A useful post hoc descriptive check is the minimum detectable effect: the effect size the planned sample would have had 80% power to detect under conventional assumptions. The purpose is to communicate what the design was capable of detecting, which is a different question from salvaging an observed p-value.

Table 16.2

Minimum detectable standardised effects under conventional two-sample testing

n per group	Rounded MDE	Interpretation
12	1.20	Only very large effects are detectable with 80% power.
20	0.91	Large effects remain the main detectable target.
50	0.57	Moderate-to-large effects become detectable.
100	0.40	Moderate effects are detectable with conventional power.

Note. Values use two-sided alpha = 0.05, 80% power, equal group sizes, and sd = 1. They should be treated as planning summaries rather than guarantees. The 100-per-group row is shown only as a comparison point and exceeds the book's primary small-sample scope.

With 12 participants per group, the design is well powered only for a standardised effect of about d = 1.20. If the smallest meaningful effect is d = 0.40, a non-significant result from that design cannot rule it out. The correct conclusion is that the study was too imprecise to decide the question.

This design-based MDE check is different from post-hoc power calculated from the observed effect size. MDE uses the planned design inputs: sample size, alpha, and a target power level. Post-hoc power based on the observed effect is circular because it is mostly a re-expression of the p-value; it does not add information after the study has been analysed and is not recommended (Hoenig and Heisey 2001).

Equivalence and Non-Inferiority

If the scientific question is whether two treatments are similar enough, ordinary null-hypothesis testing is the wrong framework. Equivalence testing starts by defining a margin: the largest difference that would still be practically negligible. The two one-sided tests procedure then asks whether the confidence interval lies entirely inside that margin (Lakens, Scheel, and Isager 2018). Non-inferiority testing uses a one-sided version when the goal is to show that a new treatment is not unacceptably worse than a standard option.

In the anxiety-score example below, the observed difference is 2 points on a 0-100 scale. The prespecified equivalence margin is +/-5 points. The ordinary p-value is non-significant, but the 95% CI extends from -3 to 7, so values above the +5 margin remain plausible.

The TOST statistics are derived from the confidence interval and margin. With df = 22, the standard error implied by the 95% CI is `SE = (7 - (-3)) / (2 * qt(0.975, 22)) = 2.41`. The lower-margin statistic is `(2 - (-5)) / SE = 2.90`; the upper-margin statistic is `(2 - 5) / SE = -1.24`. The same calculation can be reproduced with a dedicated function such as `TOSTER::TOSTtwo.raw()` when the group means, standard deviations and sample sizes are available.

```
# The values below reproduce the worked example:
# group mean difference = 2, n = 12 per group, pooled SD chosen so SE =
 ↪  2.41.
library(TOSTER)

TOSTtwo.raw(
  m1 = 52,
  m2 = 50,
  sd1 = 5.91,
  sd2 = 5.91,
  n1 = 12,
  n2 = 12,
  low_eqbound = -5,
  high_eqbound = 5,
  alpha = 0.05,
  var.equal = TRUE
)
```

When `TOSTER` is not available, the manual calculation in Table 16.3 is sufficient: compute the standard error, test the lower margin with the upper-tail probability, test the upper margin with the lower-tail probability, and conclude equivalence only if both one-sided p-values are below alpha.

Table 16.3

Two one-sided tests for the equivalence example

Test	Null	Statistic	p-value	Interpretation
Lower-margin test	Difference <= -5 points	t = 2.90	0.004	The data reject differences worse than -5 points.
Upper-margin test	Difference >= +5 points	t = -1.24	0.113	The data do not reject differences at or above +5 points.

Test	Null	Statistic	p-value	Interpretation
Overall TOST decision	Both one-sided tests must reject	–	max p = 0.113	Equivalence is not established because the upper-margin test is not significant.

Note. TOST evaluates H0: Delta <= -delta and H0: Delta >= +delta, where delta is the equivalence margin. Equivalence is concluded only if both one-sided tests reject at alpha, which is equivalent to the 100(1 - 2 alpha)% CI lying inside (-delta, +delta).

The lower-margin test rules out differences worse than -5 points, but the upper-margin test does not rule out differences of +5 points or more. The result should therefore be reported as non-significant and not equivalent. A larger or more precise study would be needed to support an equivalence claim.

Reporting Non-Significant Results

Responsible reporting avoids definitive language unless the design and interval justify it. The report should state the estimate, confidence interval, p-value, practical threshold, and design limitation. If the result is inconclusive, say so directly.

i Limitations Paragraph Template

Use a limitations paragraph to explain what the study could and could not rule out:
"This study estimated [effect/comparison] with [sample size/design]. The point estimate was [estimate] and the confidence interval ranged from [lower] to [upper]. Because the interval [does/does not] include the prespecified practically important threshold of [threshold], the result should be interpreted as [inconclusive / compatible with no meaningful effect / insufficient for equivalence]. With this sample size, the design had 80% power only for effects of approximately [MDE] or larger, so smaller effects remain unresolved. Future work should [replicate with larger n / improve measurement precision / use a paired design / pre-specify equivalence margins]."

Table 16.4

Safer language for reporting non-significant results

Weak wording	Stronger wording
There was no effect.	No statistically significant difference was observed, and the 95% CI should be used to judge plausible effects.

Weak wording	Stronger wording
The treatments were equivalent.	The observed difference was small, but the confidence interval did not stay within the prespecified equivalence margin.
The study proved the null hypothesis.	The data did not provide strong evidence against the null hypothesis; they do not prove that the effect is zero.
The result was non-significant, so smaller effects do not matter.	With n = 12 per group, the study had 80% power to detect d = 1.20; smaller effects could not be ruled out.

Note. Claims of equivalence or no meaningful effect require a prespecified margin and a sufficiently precise interval.

Bayes factors can also quantify relative evidence for a null model versus an alternative model, but they are sensitive to the prior distribution for the effect size under the alternative (Morey and Rouder 2011; Dienes 2014; Wagenmakers et al. 2018). Report the prior distribution and, when feasible, conduct a prior-sensitivity analysis showing how the Bayes factor changes across plausible specifications. In introductory small-sample reporting, a confidence-interval and equivalence-margin approach is usually more transparent for readers.

Key Takeaways

Non-significant results are not automatically negative findings. A narrow interval can rule out effects large enough to matter, whereas a wide interval leaves the study inconclusive. Equivalence and non-inferiority claims require a prespecified practical margin and evidence that the interval stays within that margin. In small-sample research, the strongest reporting pairs the p-value with the estimate, confidence interval, minimum detectable effect, and a clear statement of what remains unresolved.

Self-Assessment Quiz

Question 1

What does a non-significant result usually mean?

a) The null hypothesis is proven true
b) The data were not sufficiently inconsistent with the null hypothesis to reject it
c) The treatment has no effect
d) The study was necessarily invalid

Question 2

Which confidence interval best supports a claim of no meaningful effect if any effect within ±3 points is considered trivial?

a) -0.5 to 2.5
b) -6.0 to 6.0
c) -2.0 to 8.0
d) -10.0 to 1.0

Question 3

Why are non-significant results common in small-sample studies?

a) Small samples automatically bias effects toward zero
b) Limited power means true effects are often missed
c) Statistical tests cannot be used with small samples
d) Confidence intervals are always narrow

Question 4

What is an equivalence test designed to show?

a) That two effects are exactly identical
b) That an effect is small enough to fall within a prespecified negligible range
c) That p is greater than 0.05
d) That the sample size is large

Question 5

A non-significant result has a wide CI that includes both harmful and beneficial effects. What is the best conclusion?

a) There is no effect
b) The study is inconclusive
c) The null hypothesis is true
d) Equivalence has been shown

Question 6

Which phrase is safest when the CI is wide and p > 0.05?

a) The treatments were identical
b) No statistically significant difference was observed, and meaningful effects remain plausible
c) The intervention failed
d) The null hypothesis was confirmed

Answers and Explanations

Question 1

Answer: b)

Explanation: A non-significant result is a failure to reject the null at the chosen alpha level. It does not prove that the true effect is zero.

Why other options are not correct:

- Option a is wrong because null-hypothesis testing does not prove the null.
- Option c is wrong because an effect may exist but be estimated imprecisely.
- Option d is wrong because a valid study can still produce a non-significant result.

Question 2

Answer: a)

Explanation: The interval from -0.5 to 2.5 stays inside the practical threshold of 3 points. The wider intervals still include meaningful effects.

Why other options are not correct:

- Option b is wrong because values as large as +/-6 remain compatible with the data.
- Option c is wrong because effects above the +3-point threshold remain plausible.
- Option d is wrong because harmful effects far beyond the -3-point threshold remain plausible.

Question 3

Answer: b)

Explanation: Small samples have limited power, so they can easily produce non-significant results even when effects are real and practically meaningful.

Why other options are not correct:

- Option a is wrong because small samples increase uncertainty, not automatic bias toward zero.
- Option c is wrong because exact, rank-based and resampling tests can be used with small samples.
- Option d is wrong because small samples usually produce wider confidence intervals.

Question 4

Answer: b)

Explanation: Equivalence testing uses a practical margin and asks whether the effect is sufficiently small to be considered negligible.

Why other options are not correct:

- Option a is wrong because equivalence means close enough for the research purpose, not mathematically identical.
- Option c is wrong because $p > 0.05$ in an ordinary test does not establish equivalence.
- Option d is wrong because sample size helps precision but is not what the test is designed to show.

Question 5

Answer: b)

Explanation: A wide interval containing substantively different conclusions indicates imprecision. The appropriate conclusion is that the study is inconclusive.

Why other options are not correct:

- Option a is wrong because the interval still includes non-zero effects.
- Option c is wrong because the null is not proven by a wide non-significant interval.
- Option d is wrong because equivalence requires the interval to fit inside a prespecified negligible range.

Question 6

Answer: b)

Explanation: This wording reports the statistical result while acknowledging that the interval still includes effects that could matter.

Why other options are not correct:

- Option a is wrong because identical treatment effects are rarely shown by small-sample data.
- Option c is wrong because failure is stronger than the evidence supports when the interval is wide.
- Option d is wrong because null-hypothesis testing does not confirm the null.

Chapter 17: Transparent Reporting of Methods and Limitations

Learning Objectives

By the end of this chapter, you will be able to explain why transparent reporting is central to small-sample credibility, document analytic choices in reproducible scripts, distinguish planned from exploratory analyses, report samples, exclusions and missing data clearly, evaluate whether studies disclose enough information to support their claims, and use reporting guidelines such as CONSORT, STROBE and PRISMA as practical checklists rather than as afterthoughts.

The Importance of Transparency

Transparent reporting allows readers to evaluate the quality of evidence, assess the risk of bias, and replicate or build upon findings. With small samples, transparency is particularly important because results are more sensitive to analytic choices, outliers, and missing data. Readers need full information to judge whether conclusions are warranted.

Transparent reporting includes a clear description of sampling and recruitment, a summary of participant characteristics, complete reporting of variables and measures, a record of data cleaning and exclusions, and a justified statement of the statistical methods used. It also requires reporting all planned analyses and relevant sensitivity checks rather than only statistically significant findings. Limitations and plausible alternative explanations should be stated directly so readers can judge how far the evidence supports the conclusion.

Putting the Transparency Pieces Together

Transparency is a workflow, not a paragraph added at the end of a report. The same decisions should appear in four places: the preregistration or planning document, the analysis script, the results section and the limitations section. If those four records disagree, the report should explain why.

Stage	What the reader should be able to verify
Planning	What was primary, what was exploratory, and what decision rules were set before analysis

Stage	What the reader should be able to verify
Data preparation	How exclusions, missing data, recoding and outliers were handled
Analysis script	Which tests or models were run, with seeds, packages and sensitivity analyses visible
Results report	Estimates, intervals, p-values where relevant, adjusted p-values where needed and clear effect-size language
Limitations	How sample size, precision, design, assumptions and generalisability constrain the conclusion

This structure is especially important when the analysis changes after inspection. A change can be defensible, but it must be visible. A small-sample report should never leave readers guessing whether a method was planned, chosen because assumptions failed, or selected because it produced the strongest result.

Documenting Analytic Choices

Modern quantitative research involves many decisions: how to handle outliers, which variables to include, whether to transform variables, which test to use, how to handle missing data. These decisions, if made after seeing the data, can inflate Type I error and bias estimates (researcher degrees of freedom, p-hacking).

When possible, preregister hypotheses, methods and decision rules before data collection or before the dataset is inspected. All analysis decisions should then be documented in a reproducible script, with exploratory analyses and sensitivity checks clearly labelled. Exploratory work is entirely legitimate, but it should be labelled as such and kept separate from confirmatory analyses rather than presented as if planned from the outset.

Example: Documenting Analysis Decisions in Code Comments

A well-documented analysis script includes comments explaining each decision.

```
library(tidyverse)

# Load cleaned data (see data_cleaning.R for details)
study_data <- read_csv("data/mini_marketing.csv", show_col_types = FALSE)

# Descriptive statistics
summary(study_data)
```

```
       id           campaign   satisfaction     age_group    prior_purchase
 Min.   : 1.00   Length   :30   Min.   :1.00   Length   :30   Length   :30
 1st Qu.: 8.25   N.unique : 2   1st Qu.:3.00   N.unique : 3   N.unique : 2
 Median :15.50   N.blank  : 0   Median :4.00   N.blank  : 0   N.blank  : 0
 Mean   :15.50   Min.nchar: 5   Mean   :3.57   Min.nchar: 3   Min.nchar: 2
 3rd Qu.:22.75   Max.nchar: 6   3rd Qu.:4.00   Max.nchar: 5   Max.nchar: 3
 Max.   :30.00                  Max.   :5.00
```

```r
# Decision 1: Treat satisfaction as ordinal (1-5 scale)
# Justification: Only 5 levels; cannot assume equal intervals
# Method: Mann-Whitney U test (nonparametric)

# Decision 2: Two-sided test (no directional hypothesis preregistered)
wilcox.test(satisfaction ~ campaign, data = study_data, exact = FALSE)
```

```
    Wilcoxon rank sum test with continuity correction

data:  satisfaction by campaign
W = 156, p-value = 0.06
alternative hypothesis: true location shift is not equal to 0
```

```r
# Sensitivity analysis: Also run t-test assuming equal intervals
t.test(satisfaction ~ campaign, data = study_data, var.equal = TRUE)
```

```
    Two Sample t-test

data:  satisfaction by campaign
t = 1.8, df = 28, p-value = 0.08
alternative hypothesis: true difference in means between group Email and group Social is nc
95 percent confidence interval:
 -0.07298  1.27298
sample estimates:
 mean in group Email mean in group Social
               3.867                3.267
```

```r
# Result: Both tests yield similar p-values; conclusions robust to choice
 ↪  of test
```

Interpretation: The script documents that satisfaction is treated as ordinal and that a nonparametric test is chosen accordingly. A sensitivity analysis using a t-test (assuming equal intervals) is also reported to show robustness. This transparency helps readers understand and trust the analysis.

Describing the Sample

The sample description should state the target population, the accessible population, the sampling method, inclusion and exclusion criteria, recruitment procedures, response rate, final sample size after exclusions, and relevant participant characteristics such as demographics or baseline measures.

Use a table to summarise sample characteristics. For RCTs, report characteristics separately by group to verify balance.

Example: Sample Characteristics Table

We create a descriptive table for the `mini_marketing` dataset. The table reports group size, satisfaction scores and prior purchase rates so that readers can assess baseline comparability.

```
library(tidyverse)

# Load data
study_data <- read_csv("data/mini_marketing.csv", show_col_types = FALSE)

# Summary statistics by campaign group
summary_table <- study_data %>%
  group_by(campaign) %>%
  summarise(
    N = n(),
    `Mean Satisfaction` = round(mean(satisfaction, na.rm = TRUE), 2),
    `SD Satisfaction` = round(sd(satisfaction, na.rm = TRUE), 2),
    `Prior Purchase (%)` = round(100 * mean(prior_purchase == "Yes", na.rm
     ↪  = TRUE), 1),
    .groups = "drop"
  )

smallsamplelab_apa_table(
  "17.1",
  "Sample characteristics by campaign type",
  summary_table,
  note = "The table summarises the mini marketing study by campaign group.
   ↪  Prior purchase is reported as the percentage of participants with a
   ↪  previous purchase.",
  align = c("l", "r", "r", "r", "r"),
  col.names = c("Campaign", "n", "Satisfaction M", "Satisfaction SD",
   ↪  "Prior purchase (%)")
)
```

Table 17.1

Sample characteristics by campaign type

Campaign	n	Satisfaction M	Satisfaction SD	Prior purchase (%)
Email	15	3.87	0.99	53.3
Social	15	3.27	0.80	53.3

Note. The table summarises the mini marketing study by campaign group. Prior purchase is reported as the percentage of participants with a previous purchase.

Interpretation: The table shows sample size, satisfaction scores, and prior purchase rates for each campaign group. Readers can assess whether groups are comparable at baseline. If the study were an RCT, imbalances might suggest randomisation problems or chance variation. In observational studies, imbalances indicate potential confounding.

Reporting Missing Data

Missing-data reporting should state the number of complete observations, the number and proportion missing for each variable, visible patterns of missingness and the method used to handle missing values. If missingness clusters in certain subgroups, that pattern should be described because it may affect interpretation.

If multiple imputation was used, state the number of imputations and the imputation method.

Reporting Deviations from Planned Analyses

If the analysis plan changes after seeing the data (e.g., adding a covariate, using a different test, excluding outliers), report the deviation explicitly.

Example: "We initially planned to use a t-test but observed severe skewness in the outcome. We therefore used a Mann–Whitney U test instead. Results from both tests are reported in the supplementary materials."

Acknowledging Limitations

Every study has limitations. In small-sample studies, the common ones are limited power, wide confidence intervals, sensitivity to outliers or assumption violations, limited generalisability from narrow or non-probability samples, and inflated false-positive risk when many tests are conducted. These limitations should be connected to the interpretation: explain how they might affect the conclusion and what a future study would need to resolve.

Handling Multiple Comparisons in Small Samples

When conducting multiple statistical tests, the probability of Type I error increases. With k independent tests at $\alpha = 0.05$: - Family-wise error rate (FWER) $\approx 1 - (1 - \alpha)^k$ - For 5 tests: roughly 23% chance of at least one false positive - For 10 tests: roughly 40% chance

When to Correct

- Multiple outcomes or subgroups
- Post-hoc pairwise comparisons
- Exploratory analyses with many variables

Common Methods

1. **Bonferroni**: $\alpha_{\text{adjusted}} = \alpha / k$ (most conservative)
2. **Holm–Bonferroni**: Sequential step-down procedure
3. **Benjamini–Hochberg (FDR)**: Controls the false discovery rate

```
# Example with multiple p-values
p_values <- c(0.01, 0.03, 0.08, 0.15, 0.25)

adjustment_table <- tibble(
  Test = paste("Test", seq_along(p_values)),
  `Raw p` = p_values,
  Bonferroni = p.adjust(p_values, method = "bonferroni"),
  Holm = p.adjust(p_values, method = "holm"),
  `Benjamini-Hochberg FDR` = p.adjust(p_values, method = "fdr")
) %>%
  mutate(across(where(is.numeric), ~ sprintf("%.3f", .x)))

smallsamplelab_apa_table(
  "17.2",
  "Adjusted p-values for five exploratory tests",
  adjustment_table,
  note = "Bonferroni controls the family-wise error rate most
   ↪  conservatively; Holm is a step-down family-wise method;
   ↪  Benjamini-Hochberg controls the false discovery rate.",
  align = c("l", "r", "r", "r", "r")
)
```

Table 17.2

Adjusted p-values for five exploratory tests

Test	Raw p	Bonferroni	Holm	Benjamini–Hochberg FDR
Test 1	0.010	0.050	0.050	0.050
Test 2	0.030	0.150	0.120	0.075
Test 3	0.080	0.400	0.240	0.133
Test 4	0.150	0.750	0.300	0.188
Test 5	0.250	1.000	0.300	0.250

Note. Bonferroni controls the family-wise error rate most conservatively; Holm is a step-down family-wise method; Benjamini–Hochberg controls the false discovery rate.

Reporting Template

"We tested effects in three subgroups. After Holm–Bonferroni correction, only Group A showed a significant difference (adjusted p = 0.03)."

Small Sample Considerations

With limited power, strict corrections can remove all nominally significant findings. The practical response is to pre-specify primary outcomes, label exploratory outcomes clearly, report both corrected and uncorrected p-values where informative, and place greater weight on effect sizes and confidence intervals.

Key Takeaways

For multiple-comparison reporting, state how many tests were conducted, describe the correction method, and distinguish confirmatory from exploratory analyses. In small-sample work, adjusted p-values should usually be interpreted alongside confidence intervals because the interval shows the direction and precision of the estimate.

Pre-Registration for Small-Sample Studies

Pre-registration involves documenting your hypotheses, methods, and analysis plan before data collection begins or, at the latest, before the dataset is inspected. This is especially important for small samples because:

- Limited power increases temptation for p-hacking
- Results are more sensitive to analytic choices
- Multiple testing is common (searching for effects)
- Post-hoc storytelling is easier with small samples

What to Pre-Register

Minimum requirements: 1. Research questions and hypotheses (primary versus secondary) 2. Sample size with justification 3. Statistical tests planned for each hypothesis 4. Handling of outliers, missing data, and covariates 5. Multiple comparison corrections (if applicable)

Pre-Registration Template

Use the template below as a planning table rather than as code to run. A preregistration should be specific enough that another analyst could reproduce the planned analysis without asking what you meant.

Section	What to write before analysis
Study title	Short descriptive title and date of preregistration
Primary question	One confirmatory question stated in testable terms
Secondary questions	Exploratory or supportive questions labelled as secondary
Hypotheses	Directional or non-directional predictions, including the expected outcome metric
Design and sample	Target sample size, stopping rule, recruitment source, inclusion and exclusion criteria
Variables	Primary outcome, predictors, covariates and scoring rules
Primary analysis	Statistical test or model, alpha level, effect size, confidence interval and software
Assumption checks	Planned diagnostics and what will be done if assumptions are not met
Outliers and missing data	Definitions, handling rules and sensitivity analyses
Multiplicity	Which outcomes are primary, which are exploratory and how p-values will be adjusted
Decision rule	What pattern of estimate, interval and p-value will be interpreted as support for the primary hypothesis

Where to Pre-Register

The **Open Science Framework** (osf.io) provides free, time-stamped registration for study protocols, analysis plans and materials. **AsPredicted** (aspredicted.org) provides a short nine-question

template that is widely used for behavioural, psychology and management studies. **Registered Reports** are a journal submission format in which the research question and methods are reviewed before results are known, with in-principle acceptance if the protocol is judged sound.

Handling Deviations

Deviations are acceptable if reported transparently:

```
**Deviations from Pre-Registration:**
1. Sample size: Planned n = 40, achieved n = 36 due to [reason]
2. Primary test: Switched from t-test to Mann-Whitney due to severe
 ↪  skewness (skew = 2.4)
3. Additional analysis: Added baseline covariate per reviewer request
 ↪  (post-hoc, clearly labelled)
```

Benefits for Small Samples

- Protects against p-hacking accusations
- Separates confirmatory from exploratory analyses
- Improves study design through upfront planning
- Facilitates transparent reporting

Pre-Registration Checklist

☐ Hypotheses specific and testable
☐ Sample size justified
☐ All variables operationally defined

☐ Statistical tests specified
☐ Outlier/missing data plans stated
☐ Multiple comparison approach stated
☐ Time-stamped before analysis

Following Reporting Guidelines

Numerous reporting guidelines exist for different study designs:

- **CONSORT** (Schulz, Altman, and Moher 2010): Randomised controlled trials.
- **STROBE** (Elm et al. 2007): Observational studies (cohort, case-control, cross-sectional).
- **PRISMA** (Page et al. 2021): Systematic reviews and meta-analyses.
- **COREQ**: Qualitative research.

These guidelines provide checklists of items to report. Following them improves transparency and comparability across studies. Even if formal adherence is not required, consult the relevant guideline as a checklist.

For example, CONSORT asks randomised trials to report participant flow. In a small pilot RCT, this can be as simple as a table that states how many participants were assessed, randomised, analysed and excluded at each stage. If 40 people were screened, 30 were enrolled, and 28 were analysed, the report should make clear where the two losses occurred and whether they were related to group assignment or outcome.

Table 17.3

Example participant-flow summary for a small pilot RCT

Stage	n	Note
Assessed for eligibility	40	10 did not meet inclusion criteria or declined
Randomised	30	1:1 allocation
Allocated to intervention	15	14 analysed; 1 withdrew before post-test
Allocated to control	15	14 analysed; 1 missing post-test
Included in analysis	28	Primary analysis used available paired outcomes
Excluded after randomisation	2	Reasons reported by group

Note. This table illustrates the reporting logic of CONSORT Item 13a. A full trial report would usually include a flow diagram as well.

Key Takeaways

Transparent reporting allows readers to evaluate the quality and limits of small-sample evidence. The essential tasks are to document analytic choices in reproducible scripts, report sample characteristics, missing data and exclusions explicitly, disclose deviations from planned analyses, and present sensitivity analyses where decisions could affect the result. Relevant reporting guidelines such as CONSORT and STROBE should be used as checklists, while the limitations section should state clearly how sample size, precision, assumptions and generalisability affect the conclusion.

Self-Assessment Quiz

Test your understanding of transparent reporting from Chapter 17.

Question 1

Which should be reported when documenting a small-sample study?

a) Only significant results
b) All analyses conducted, including non-significant findings
c) Only the primary analysis
d) Results can be selectively reported

Question 2

A study planned to use a t-test but switched to Mann–Whitney after seeing skewed data. How should this be reported?

a) Do not mention the change
b) Report only the Mann–Whitney result
c) State the planned test, explain the skewness, report the Mann–Whitney result and include a sensitivity analysis
d) Pretend Mann–Whitney was always planned

Question 3

What is “p-hacking”?

a) Illegally accessing data
b) Trying multiple analyses/subgroups until finding $p<0.05$, then reporting only that result
c) Using permutation tests
d) Adjusting for multiple comparisons

Question 4

Pre-registration helps prevent:

a) Sample size limitations
b) Researcher degrees of freedom (flexibility in analysis choices) leading to false positives
c) Missing data
d) Measurement error

Question 5

A study with n=15 per group finds p=0.12. The limitation section should state:

a) "The result is not significant, proving no effect exists"
b) "The study was underpowered to detect small-to-medium effects; findings are inconclusive"
c) "The sample size was adequate"
d) Nothing; non-significant results need no discussion

Answers and Explanations

Question 1

Answer: b)

Explanation: Transparent reporting requires documenting planned analyses, exploratory analyses and sensitivity checks, not just significant findings. Selective reporting inflates Type I error across the literature and prevents readers from evaluating the quality of the evidence.

Question 2

Answer: c)

Explanation: Deviations from plans should be documented with justification. Reporting both the planned and adapted analyses shows how much the conclusion depends on the analytic choice.

Question 3

Answer: b)

Explanation: P-hacking involves exploring many analyses, such as different covariates, subgroups or outlier rules, until statistical significance appears and then selectively reporting that analysis. This inflates the false-positive rate.

Question 4

Answer: b)

Explanation: Pre-registration documents hypotheses and analysis plans before the data are inspected. This reduces post-hoc decisions that capitalise on chance and inflate Type I error.

Question 5

Answer: b)

Explanation: Small samples have limited power. Non-significance may reflect insufficient power rather than absence of effect, so the limitation section should discuss precision, minimum detectable effects and the uncertainty around the estimate.

Chapter 18: Visualising Uncertainty and Presenting Results

Learning Objectives

By the end of this chapter, you will be able to explain why uncertainty visualisation is central to small-sample reporting, distinguish standard deviations, standard errors and confidence intervals, show individual observations alongside summaries, create clear `ggplot2` figures with uncertainty intervals, identify misleading visual choices, and design figures and tables that support estimation rather than binary significance claims.

The Role of Visualisation in Small-Sample Research

Visualisation serves multiple purposes:

- **Exploratory**: Identify patterns, outliers, and distributional features during data screening.
- **Diagnostic**: Assess assumptions (normality, linearity, homoscedasticity).
- **Inferential**: Display estimates, confidence intervals, and group comparisons.
- **Communicative**: Convey findings to diverse audiences in accessible formats.

With small samples, visualisation is particularly valuable because individual data points can be shown (unlike large datasets where summaries are necessary). Showing raw data alongside summaries makes the variability and structure of the data visible to the reader.

Visualising Point Estimates with Confidence Intervals

Error bars (standard errors or confidence intervals) convey uncertainty. Use 95% CIs for inferential plots, as they align with conventional significance testing (CIs that exclude zero correspond to $p < 0.05$).

Best practices: - Label axes clearly with units. - Include a legend if multiple groups are compared. - Use colour or shape to distinguish groups. - Avoid 3D effects and unnecessary decoration (chart junk). - Use colour-blind-safe palettes such as `viridis` or carefully chosen ColorBrewer palettes. Do not rely on colour alone. Combine colour with labels, shapes, or direct annotation where possible.

Example: Bar Plot with Error Bars

We compare mean satisfaction scores between two campaign types with 95% CI error bars.

```r
library(tidyverse)

# Load data
study_data <- read_csv("data/mini_marketing.csv", show_col_types = FALSE)

# Compute means and 95% CIs
summary_stats <- study_data %>%
  group_by(campaign) %>%
  summarise(
    mean_satisfaction = mean(satisfaction, na.rm = TRUE),
    se = sd(satisfaction, na.rm = TRUE) / sqrt(n()),
    t_crit = qt(0.975, df = n() - 1),
    ci_lower = mean_satisfaction - t_crit * se,
    ci_upper = mean_satisfaction + t_crit * se,
    .groups = "drop"
  )

# Bar plot with error bars
ggplot(summary_stats, aes(x = campaign, y = mean_satisfaction, fill =
 ↪  campaign)) +
  geom_col(width = 0.6, alpha = 0.7) +
  geom_errorbar(aes(ymin = ci_lower, ymax = ci_upper), width = 0.2) +
  labs(
    title = "Mean Customer Satisfaction by Campaign Type",
    x = "Campaign",
    y = "Satisfaction (1-5 scale)",
    fill = "Campaign"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```

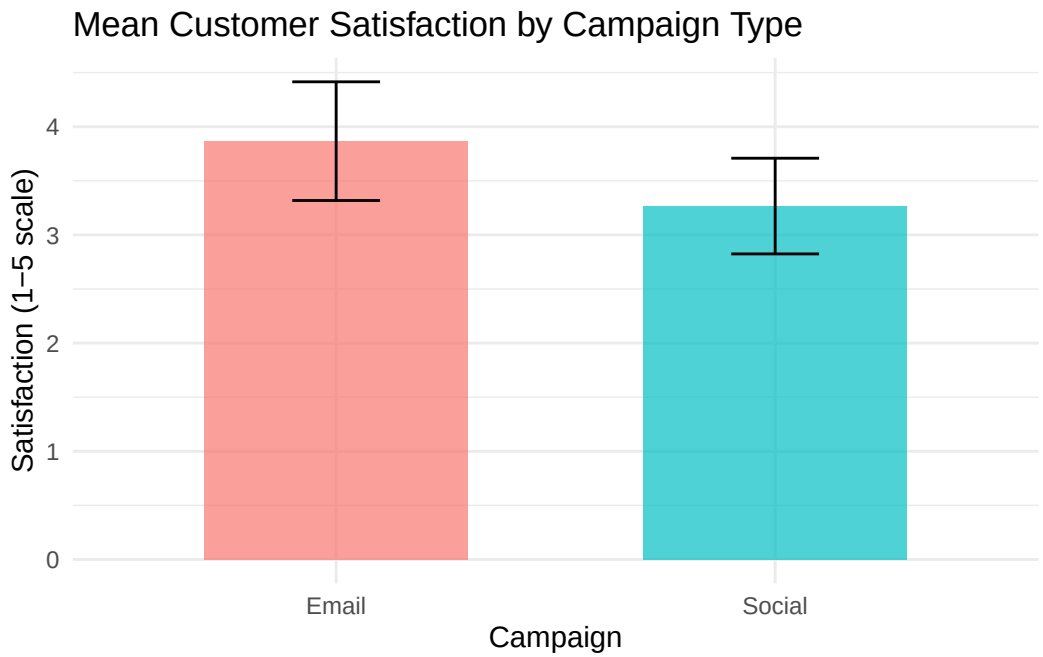


Figure 18.1: Mean customer satisfaction by campaign type with 95% confidence intervals.

Interpretation: The bars show mean satisfaction for each campaign, and the error bars show 95% confidence intervals based on the t distribution within each group. These intervals describe uncertainty around each group mean. Formal group comparisons should still be reported with the planned statistical test rather than judged only by overlap of the bars.

Showing Individual Data Points

With small samples (n < 50), individual data points can be overlaid on summary plots. This reveals the distribution, identifies outliers, and shows sample size directly.

Example: Dot Plot with Mean and CI

We create a dot plot showing individual satisfaction scores, overlaid with group means and CIs.

```
library(tidyverse)

study_data <- read_csv("data/mini_marketing.csv", show_col_types = FALSE)

# Compute summary statistics
summary_stats <- study_data %>%
  group_by(campaign) %>%
```

```r
  summarise(
    mean_satisfaction = mean(satisfaction, na.rm = TRUE),
    se = sd(satisfaction, na.rm = TRUE) / sqrt(n()),
    t_crit = qt(0.975, df = n() - 1),
    ci_lower = mean_satisfaction - t_crit * se,
    ci_upper = mean_satisfaction + t_crit * se,
    .groups = "drop"
  )

# Dot plot with mean and CI
ggplot(study_data, aes(x = campaign, y = satisfaction, colour = campaign))
 ↪  +
  geom_jitter(width = 0.1, alpha = 0.6, size = 2) +
  geom_point(data = summary_stats, aes(y = mean_satisfaction),
             size = 4, shape = 18, colour = "black") +
  geom_errorbar(data = summary_stats, aes(y = mean_satisfaction, ymin =
   ↪  ci_lower, ymax = ci_upper),
                width = 0.2, colour = "black", linewidth = 1) +
  labs(
    title = "Customer Satisfaction by Campaign Type",
    subtitle = "Individual scores (dots), mean (diamond), and 95% CI
     ↪  (error bars)",
    x = "Campaign",
    y = "Satisfaction (1-5 scale)"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```

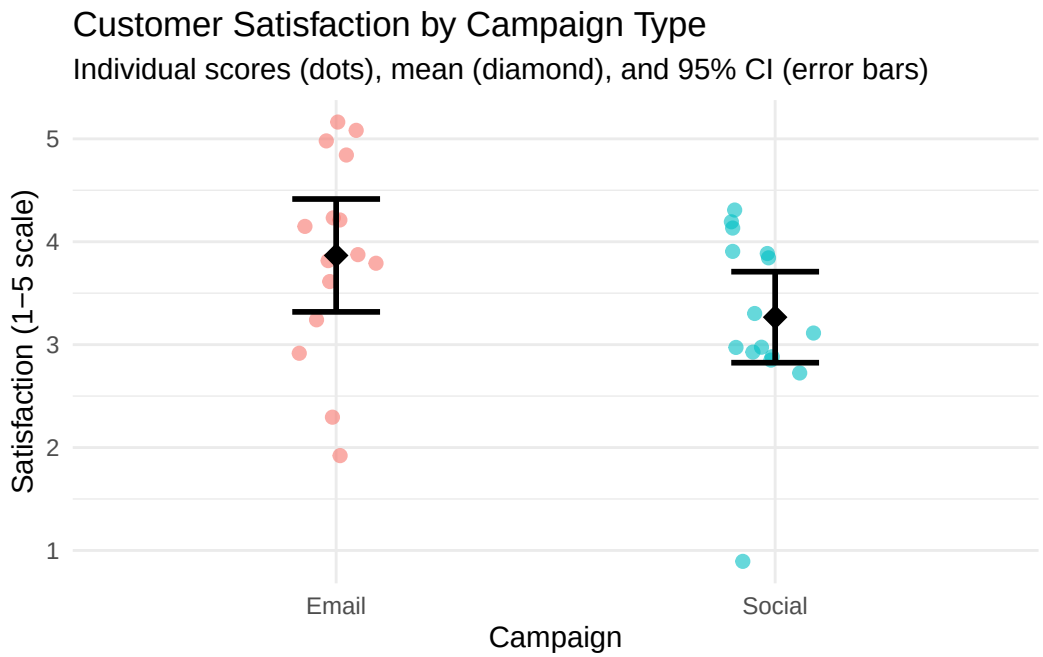


Figure 18.2: Individual satisfaction scores with group means and 95% confidence intervals.

Interpretation: Each dot represents one participant. The diamond shows the group mean, and the error bars show the 95% CI. Readers can see the distribution of individual scores, the central tendency, and the precision of the estimate simultaneously.

Box Plots for Distributional Comparison

Box plots display the median, quartiles, and outliers, providing a non-parametric summary of distribution. They are particularly useful for comparing groups when data are skewed or ordinal.

Example: Box Plot Comparison

We create a box plot comparing satisfaction scores between campaigns.

```
library(tidyverse)

study_data <- read_csv("data/mini_marketing.csv", show_col_types = FALSE)

ggplot(study_data, aes(x = campaign, y = satisfaction, fill = campaign)) +
  geom_boxplot(alpha = 0.7) +
  geom_jitter(width = 0.1, alpha = 0.4) +
  labs(
    title = "Customer Satisfaction by Campaign Type",
```

```
    subtitle = "Box plot with individual data points",
    x = "Campaign",
    y = "Satisfaction (1-5 scale)"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```

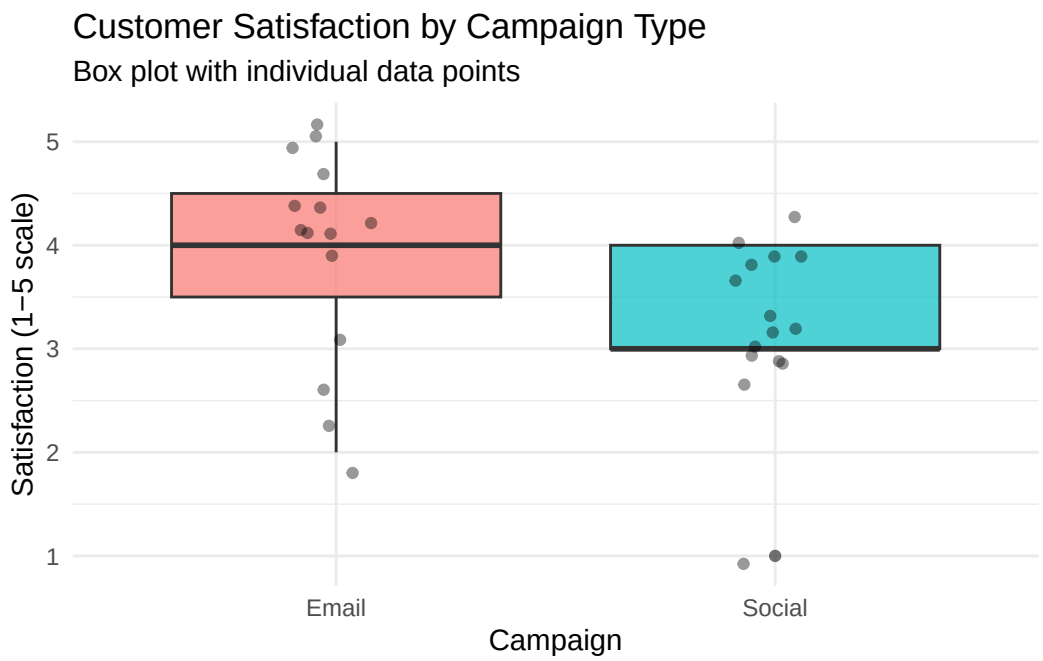


Figure 18.3: Box plot comparing customer satisfaction across campaign types.

Interpretation: The box shows the interquartile range (IQR) with the median as a line inside. Whiskers extend to 1.5 × IQR, and points beyond are potential outliers. Overlaying individual points shows sample size and exact values. This plot is ideal for nonparametric comparisons, such as the Mann–Whitney U test.

Visualising Regression Results

For regression models, plot predicted values with confidence bands, and overlay observed data. This shows model fit, uncertainty, and deviations.

Example: Scatterplot with Regression Line and CI Band

We fit a linear regression (performance ~ experience) and plot the results.

```r
library(tidyverse)

set.seed(2025)

# Simulated data
reg_data <- tibble(
  experience = runif(20, 1, 10),
  performance = 50 + 3 * experience + rnorm(20, 0, 5)
)

# Fit model
model <- lm(performance ~ experience, data = reg_data)

# Plot with regression line and CI band
ggplot(reg_data, aes(x = experience, y = performance)) +
  geom_point(size = 3, alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, colour = "blue", fill =
  ↪  "lightblue") +
  labs(
    title = "Performance vs. Experience",
    subtitle = "Linear regression with 95% confidence band",
    x = "Years of Experience",
    y = "Performance Score"
  ) +
  theme_minimal()
```

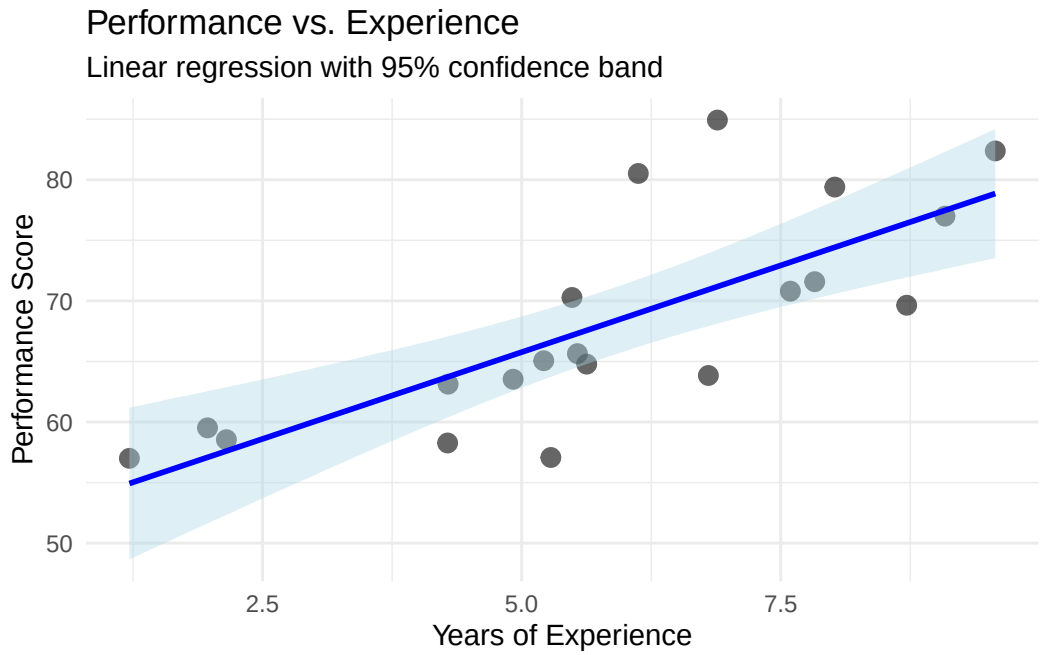


Figure 18.4: Linear regression of performance on experience with 95% confidence band.

Interpretation: Each point is an observed case. The blue line is the fitted regression line. The shaded band is the 95% confidence interval for the predicted mean at each value of experience. The band widens at the extremes (where data are sparse), reflecting greater uncertainty. This visualisation shows model fit, precision, and individual deviations simultaneously.

Forest Plots for Several Estimates

When a report compares several estimates, a forest plot is often clearer than several separate tables. The plot should show the point estimate, the confidence interval, and a reference line such as zero for mean differences or one for ratios. This format works well for multiple outcomes, subgroup estimates, or sensitivity analyses.

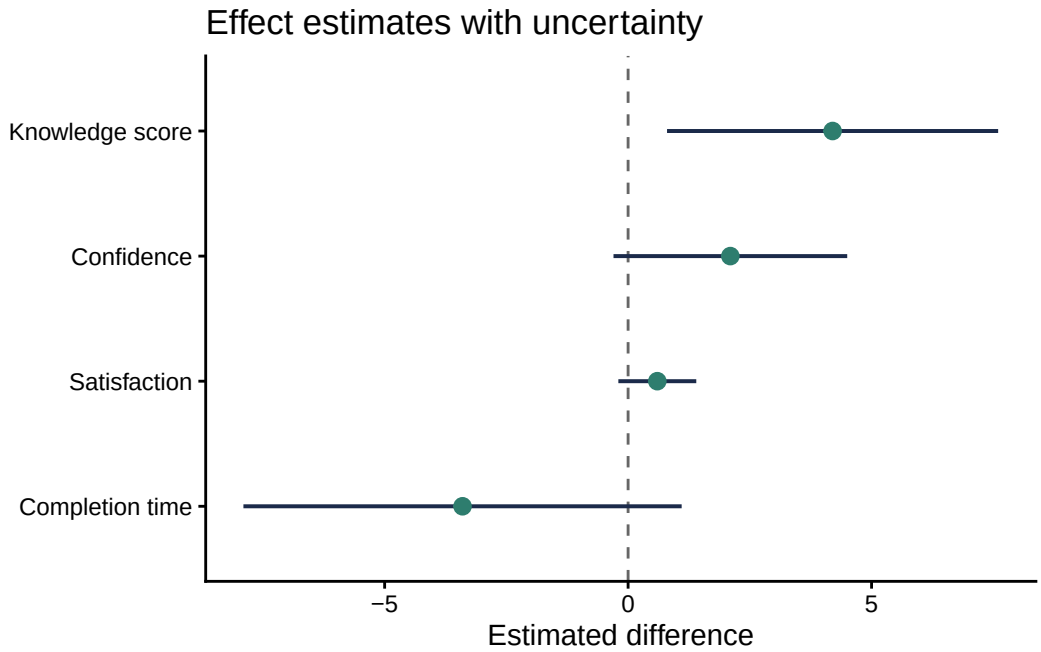


Figure 18.5: Forest plot of four small-sample estimates with 95% confidence intervals.

Interpretation: The forest plot makes precision visible. The knowledge-score interval stays above zero, while the other intervals include zero or values close to it. This does not make the other estimates unimportant. It shows that the small sample leaves more uncertainty about their direction or practical importance.

Raincloud and Half-Eye Plots

Raincloud or half-eye plots combine the raw observations, a distribution summary and an interval display. They are useful when a small sample is large enough to show distributional shape but small enough that individual observations should remain visible. The `ggdist` package provides a compact implementation. Because it is an optional visualisation package, the example below is shown as a template.

```
library(ggdist)

ggplot(study_data, aes(x = satisfaction, y = campaign, fill = campaign)) +
  ggdist::stat_halfeye(adjust = 0.8, width = 0.55, alpha = 0.6) +
  geom_jitter(height = 0.08, width = 0, alpha = 0.6, size = 1.8) +
  stat_summary(fun = median, geom = "point", shape = 23, size = 2.6, fill
  ↪  = "white") +
  scale_fill_viridis_d(option = "C", end = 0.8) +
  labs(x = "Satisfaction (1-5 scale)", y = "Campaign") +
  theme_classic(base_size = 11) +
```

```
  theme(legend.position = "none")
```

Presenting Results in Tables

Tables complement figures by providing exact values. For small samples, consider showing:

- Sample sizes (n per group).
- Means and standard deviations (or medians and IQRs).
- Effect sizes and confidence intervals.
- Test statistics and p-values.

Use simple publication tables with clear labels, sample sizes and notes that explain the inferential method.

For publication tables, align text columns left and numeric columns right, use a consistent number of decimal places within a column, and put method details in notes rather than in crowded column headings. Report exact sample sizes, avoid unnecessary trailing precision, and state whether intervals are exact, bootstrap, model-based or rank-based. If a table mixes estimates with different scales, use separate sections or clear row labels so readers do not compare incompatible numbers.

Example: Results Summary Table

We create a summary table for the campaign comparison.

```
library(tidyverse)
study_data <- read_csv("data/mini_marketing.csv", show_col_types = FALSE)

# Summary statistics
summary_table <- study_data %>%
  group_by(campaign) %>%
  summarise(
    N = n(),
    Mean = round(mean(satisfaction, na.rm = TRUE), 2),
    SD = round(sd(satisfaction, na.rm = TRUE), 2),
    Median = median(satisfaction, na.rm = TRUE),
    .groups = "drop"
  )

# Mann-Whitney test
mw_result <- wilcox.test(satisfaction ~ campaign, data = study_data, exact
 ↪  = FALSE)
p_value <- mw_result$p.value
test_note <- if (p_value < 0.001) {
```

```
  "Mann-Whitney rank-sum test: p < 0.001."
} else {
  sprintf("Mann-Whitney rank-sum test: p = %.3f.", p_value)
}

smallsamplelab_apa_table(
  "18.1",
  "Summary statistics and Mann-Whitney test for campaign satisfaction",
  summary_table,
  note = test_note,
  align = c("l", "r", "r", "r", "r")
)
```

Table 18.1

Summary statistics and Mann–Whitney test for campaign satisfaction

Table 0.1

campaign	N	Mean	SD	Median
Email	15	3.87	0.99	4
Social	15	3.27	0.80	3

Note. Mann–Whitney rank-sum test: p = 0.057.

Interpretation: The table provides exact summary statistics for each group. Readers can see sample sizes, central tendency, and variability. The p-value from the Mann–Whitney test is reported in the subtitle or a footnote. Tables and figures together provide a complete, accessible presentation of results.

Avoiding Misleading Visualisations

Common pitfalls:

- **Suppressed zero on the y-axis**: Exaggerates differences. Use a zero baseline unless there is good reason not to (and explain the choice).
- **3D effects and unnecessary decoration**: Distract from data and can obscure values.
- **Dual axes with different scales**: Misleading comparisons. Avoid or use with extreme caution.
- **Overplotting without jitter or transparency**: Hides overlapping points. Use jitter, transparency, or both.

Key Takeaways

Small-sample figures should show uncertainty and, whenever feasible, the individual observations behind the summary. Point estimates should be paired with confidence intervals, box plots and dot plots should reveal distributional features, and regression figures should show both fitted trends and uncertainty bands. Tables remain necessary because they provide exact sample sizes, estimates, intervals and test results. The central principle is transparency: avoid suppressed axes, 3D effects, dual axes and other design choices that make modest evidence appear stronger than it is.

Self-Assessment Quiz

Question 1

Why is visualisation particularly valuable in small-sample research?

a) Large datasets cannot be visualised
b) Individual data points can be shown alongside summaries, revealing variability and building trust
c) Visualisation eliminates the need for statistical tests
d) Small samples require 3D plots

Question 2

What do 95% confidence interval error bars represent?

a) The range containing all data points
b) The uncertainty around a point estimate
c) The standard deviation
d) The sample size

Question 3

What is the advantage of overlaying individual data points on summary plots (means or medians)?

a) It makes the plot more colourful
b) It reveals the distribution, identifies outliers, and shows sample size directly
c) It is required by journal guidelines
d) It eliminates the need for error bars

Question 4

In a regression plot with a confidence band, why does the band typically widen at the extremes?

a) It is a plotting error
b) The band widens where data are sparse, reflecting greater uncertainty in predictions
c) The confidence band is always the same width
d) It indicates the sample size

Question 5

What information does a box plot display?

a) Only the mean
b) The median, quartiles (IQR), and outliers
c) Individual data points only
d) The correlation coefficient

Question 6

Why should suppressed zero on the y-axis be avoided (or used with caution)?

a) It saves space
b) It can exaggerate differences and mislead readers about the magnitude of effects
c) It is required by statistical standards
d) It improves visual appeal

Question 7

What is “chart junk”?

a) High-quality graphics
b) Unnecessary decoration (3D effects, excessive colors, ornaments) that distracts from data
c) Statistical error bars
d) Individual data points

Question 8

When creating results tables, what essential information should be included?

a) Only p-values
b) Sample sizes, means/medians, measures of variability (SD or IQR), confidence intervals, and test results
c) Only the mean values
d) Just the hypothesis

Question 9

What does jittering accomplish in plots with many overlapping points?

a) It removes outliers
b) It adds small random offsets to points to reveal overlapping observations
c) It changes the statistical significance
d) It increases sample size

Question 10

In the dot plot example showing satisfaction scores, what does the diamond symbol represent?

a) An outlier
b) The group mean
c) A missing value
d) The maximum value

Answers and Explanations

Question 1

Answer: b)

Explanation: Small samples make it practical to show individual observations alongside summaries. This lets readers see variability, outliers and sample size directly rather than relying only on means or p-values.

Question 2

Answer: b)

Explanation: Confidence intervals show uncertainty around an estimate, such as a mean, difference, rate or regression coefficient.

Question 3

Answer: b)

Explanation: Showing raw data reveals patterns that summary statistics alone can hide, including skewness, ties, clusters and influential observations.

Question 4

Answer: b)

Explanation: Confidence bands usually widen where there are fewer observations to constrain the model, often near the edges of the predictor range.

Question 5

Answer: b)

Explanation: A box plot shows the interquartile range, the median and potential outliers. It is a compact non-parametric summary of a distribution.

Question 6

Answer: b)

Explanation: Starting an axis above zero can make small differences appear larger than they are. If a truncated axis is necessary, the reason should be stated clearly.

Question 7

Answer: b)

Explanation: Chart junk refers to non-data elements that reduce clarity, such as unnecessary 3D effects, excessive colour or decorative elements.

Question 8

Answer: b)

Explanation: Complete tables report sample size, central tendency, variability, effect size, interval estimates and the planned test result so readers can assess both statistical and practical importance.

Question 9

Answer: b)

Explanation: Jittering adds small random offsets to points so overlapping observations become visible.

Question 10

Answer: b)

Explanation: The diamond marks the group mean, while the smaller dots show individual observations. Different symbols help distinguish summaries from raw data.

Summary of Part D

Part D focused on how to report small-sample findings clearly and responsibly. These chapters emphasised effect sizes and confidence intervals over isolated p-values, careful interpretation of non-significant results, transparent reporting of methods and limitations, and figures and tables that display uncertainty rather than hiding it.

The shared principle is that reporting is part of the analysis, not a cosmetic final step. With small samples, readers need clear descriptions of the design, the analytic choices, the uncertainty and the limits of the evidence in order to judge what the study actually shows. A well-reported small study does not overstate its findings. It makes the available information interpretable.

Part E: Worked Projects

This part presents complete case studies that integrate methods from earlier chapters. Each project includes background, research questions, data description, analysis with code and interpretation, sensitivity analyses, visualisations, and a discussion of findings and limitations. These examples demonstrate how to combine multiple techniques to address realistic small-sample research problems.

In This Part

- Project 1. Evaluating a Marketing Campaign with Ordinal Outcomes
- Project 2. Assessing Reliability of a Short Service Quality Scale
- Project 3. Evaluating a Process Improvement Intervention (Paired Design)
- Project 4. Evaluating a Reading Intervention in Small Classrooms (Education)
- Project 5. Understanding Intervention Mechanisms: Simple Mediation Analysis
- Summary of Part E

Project 1. Evaluating a Marketing Campaign with Ordinal Outcomes

Background

A small marketing team tested two campaign formats, email and social media, with 15 customers in each group. The outcome is a 1-5 satisfaction rating, so the primary analysis should respect the ordinal scale and the small sample size rather than relying only on a normal-theory mean comparison.

Research Question

The practical question is whether the email campaign tends to produce higher customer satisfaction than the social media campaign. Because the sample is small and the ratings contain ties, the Mann–Whitney rank-sum test is a defensible primary analysis, supported by descriptive summaries and a sensitivity check using the original score scale.

Descriptive Summary

Table P1.1

Satisfaction ratings by campaign

Campaign	n	Median	IQR	Mean	SD
Email	15	4.0	1.0	3.87	0.99
Social	15	3.0	1.0	3.27	0.80

Note. Ratings are ordinal responses from 1 (low satisfaction) to 5 (high satisfaction).

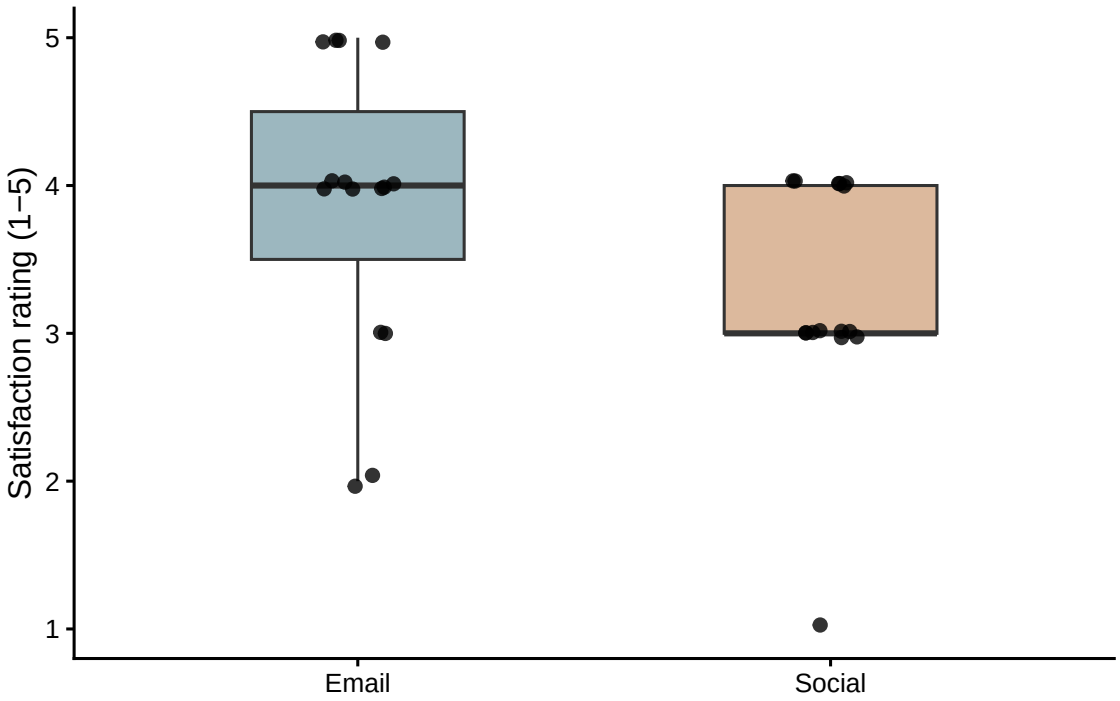


Figure P1.1: Distribution of satisfaction ratings by campaign.

The email group has a higher median and mean rating, but the distributions overlap. That overlap is important: in a small ordinal study, the result should be reported as evidence of a tendency rather than as a definitive campaign ranking.

Primary Analysis

Table P1.2

Mann–Whitney rank-sum analysis for campaign satisfaction

Quantity	Value
Wilcoxon rank-sum W	156
Two-sided p-value	0.057
Hodges–Lehmann shift	1.00
95% CI for shift	-0.00 to 1.00
Cliff's delta	0.39

Note. The test uses a normal approximation because tied ordinal scores make the exact rank-sum distribution less straightforward.

The Hodges–Lehmann shift is about one rating point in favour of email, but the confidence interval reaches the null boundary. The result is better described as suggestive than conclusive. Cliff's delta is positive, indicating that a randomly selected email respondent tends to rate satisfaction higher than a randomly selected social-media respondent, but the small sample limits precision.

Sensitivity and Subgroup Checks

Table P1.3

Primary and sensitivity analyses for the campaign comparison

Analysis	Estimate	95% CI	p-value
Rank-sum primary analysis	Shift = 1.00	-0.00 to 1.00	0.057
Equal-variance t-test sensitivity	Mean difference = 0.60	-0.07 to 1.27	0.078

Note. The t-test is a sensitivity analysis on the numeric rating scale, not the primary analysis for the ordinal outcome.

Table P1.4

Campaign satisfaction by prior-purchase history

Campaign	Prior purchase	n	Median	Mean
Email	No	7	4.0	3.86
Email	Yes	8	4.0	3.88
Social	No	7	4.0	3.29
Social	Yes	8	3.0	3.25

Note. Cells are small, so these subgroup summaries are descriptive only.

The sensitivity analysis points in the same direction as the rank-based analysis. The subgroup table should not be used to claim moderation because several cells contain only a few customers. Its role is diagnostic: it checks whether the campaign comparison is being driven by a visibly uneven customer mix.

Reporting Summary

In this mini-study, the email campaign produced higher observed satisfaction ratings than the social-media campaign. The rank-sum test was borderline, W = 156, p = 0.057, with an estimated location shift of about 1.00 rating point. A cautious report would say that the result is consistent with a modest email advantage, but that the sample is too small to treat the campaign difference as settled.

Extension Task

Re-run the analysis after treating satisfaction as a binary outcome, for example ratings of 4 or 5 versus lower ratings. Compare the Fisher exact test result with the rank-sum result and write two sentences explaining what information is lost when the ordinal scale is collapsed.

Project 2. Assessing Reliability of a Short Service Quality Scale

Background

A customer-service team piloted a three-item service-quality scale with 36 respondents across three branches. The aim is to assess whether the items are coherent enough for a pilot composite and to identify any wording or distribution problems before the study is scaled up.

Item Distributions

Table P2.1

Item descriptive statistics for the service-quality pilot

Item	Mean	SD	Median	Floor %	Ceiling %
Clarity	5.08	1.38	5.0	2.8	16.7
Professionalism	4.69	1.43	5.0	8.3	11.1
Responsiveness	4.61	1.34	4.0	5.6	11.1

Note. Floor percentage is the proportion of respondents scoring at the lowest possible scale value; ceiling percentage is the proportion scoring at the highest possible scale value.

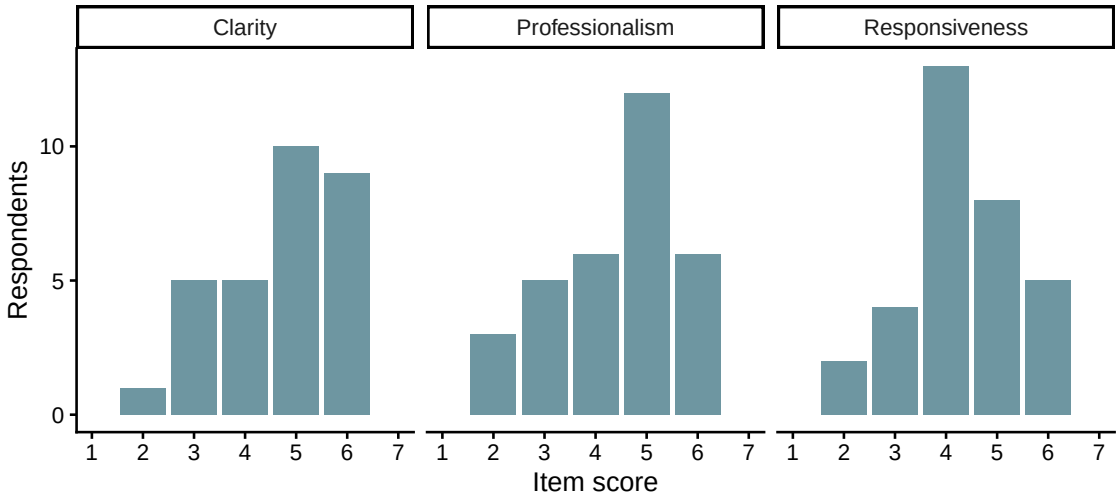


Figure P2.1: Item score distributions for the three-item service-quality scale.

The item distributions do not show severe floor or ceiling compression. With only three items, the most important checks are conceptual coherence, item-total correlations, and the width of the reliability interval.

Reliability Summary

Table P2.2

Internal-consistency summary for the short scale

Quantity	Value
Cronbach's alpha	0.793
Bootstrap 95% CI	0.633 to 0.884
Mean inter-item correlation	0.564
Number of items	3
Sample size	36

Note. The confidence interval uses the asymptotic standard error reported by psych::alpha(); with n = 36 it should be treated as approximate.

The alpha estimate is acceptable for a short pilot scale, and the mean inter-item correlation is within a plausible range for a narrow construct. This does not prove unidimensionality, but it supports using a provisional composite while documenting uncertainty.

Item Diagnostics

Table P2.3

Corrected item-total diagnostics

Item	Corrected item-total r	Alpha if deleted	Decision
Responsiveness	0.719	0.631	Retain
Professionalism	0.593	0.766	Retain
Clarity	0.601	0.755	Retain

Note. Values below about 0.30 would usually trigger a wording and content review rather than automatic deletion.

All three corrected item-total correlations exceed 0.30. The deletion diagnostics do not suggest that a single item is undermining the scale, so the next revision step should focus on respondent feedback and content coverage rather than statistical item removal.

Branch-Level Check

Table P2.4

Reliability and composite means by branch

Branch	n	Alpha	Mean composite
East	8	0.769	4.54
North	15	0.578	5.07
South	13	0.886	4.64

Note. Branch-level alpha estimates are descriptive because each subgroup is very small.

The branch summaries are useful for screening but not for formal comparison. If one branch showed a markedly different alpha or composite mean, the appropriate response would be to inspect administration conditions and item interpretation, not to claim a branch-level psychometric difference.

Reporting Summary

The three-item pilot scale showed acceptable internal consistency, alpha = 0.793, bootstrap 95% CI [0.633, 0.884]. Corrected item-total correlations were all above 0.30. The scale can be used as a provisional pilot composite, but a larger sample and qualitative item review are still needed before treating it as an established measure.

Extension Task

Create a revised two-item version by removing one item of your choice, then recompute alpha, the mean inter-item correlation and the composite score. Write a short justification for whether the shorter version improves interpretability or merely reduces content coverage.

Project 3. Evaluating a Process Improvement Intervention (Paired Design)

Background

A small operations team tracked incident counts for 20 units before and after a process change. Because each unit is measured twice, the analysis should use the paired structure. Lower post-intervention counts are favourable, so the practical estimand is the within-unit reduction.

Descriptive Summary

Table P3.1

Before-after incident summaries

Time	Mean	SD	Median
Before	4.30	1.56	4.5
After	2.75	1.65	3.0
Reduction (before - after)	1.55	2.24	2.0

Note. Reduction is defined as before minus after; positive values indicate fewer incidents after the change.

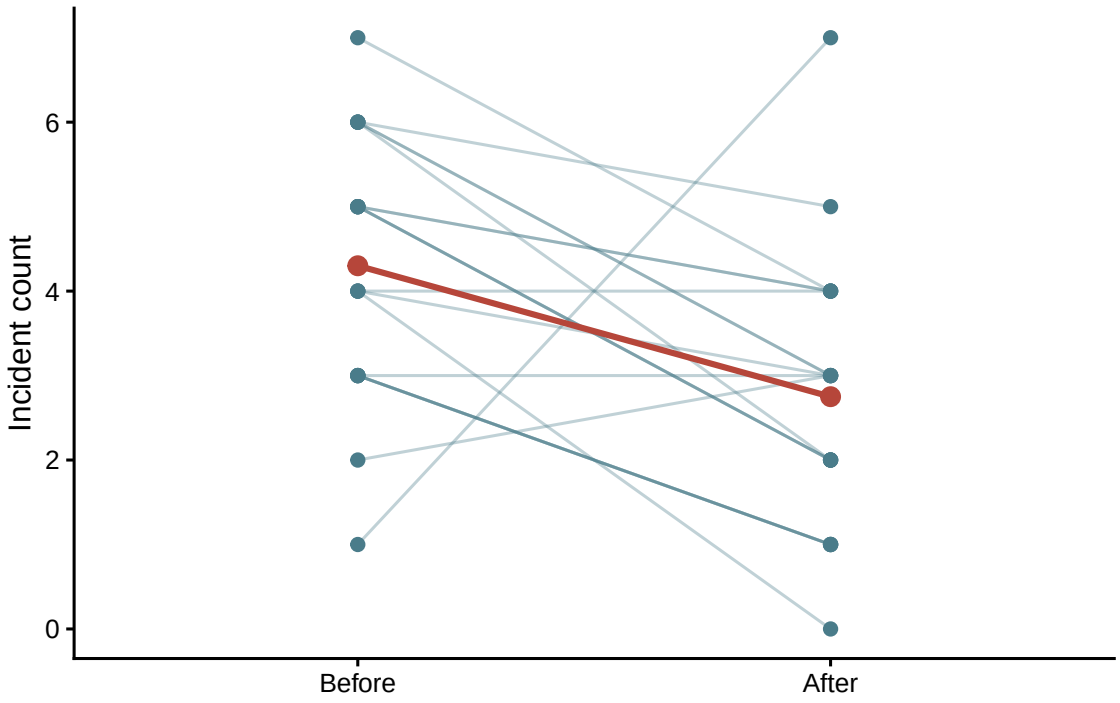


Figure P3.1: Within-unit incident counts before and after the process change.

Most units show lower incident counts after the process change, although the individual trajectories are not identical. The connected-line display is more informative than two independent boxplots because it shows the paired changes directly.

Primary and Sensitivity Analyses

Table P3.2

Paired analysis of the process change

Analysis	Estimate	95% CI	p-value
Paired t-test	Mean after-before = -1.55	-2.60 to -0.50	0.006
Wilcoxon signed-rank	Pseudomedian after-before = -2.00	-2.50 to -1.00	0.005
Paired standardised mean change	dz = -0.69	Not computed	–

Note. The signed-rank test is a robustness check for skewed paired differences. Negative after-before estimates indicate improvement because lower counts are better; dz is computed as after minus before, so the negative sign denotes a standardised reduction.

The paired t-test and signed-rank test agree that incident counts were lower after the process change. The mean reduction is about 1.55 incidents per unit. With n = 20, the evidence is strong for this dataset, but the estimate should still be reported with its confidence interval rather than only as a p-value.

Diagnostic Checks

Table P3.3

IQR-screened unusual paired changes

Unit	Department	Before	After	Reduction	Flag
1	QA	1	7	-6	Review

Note. Outlier screening identifies units for audit; it is not an automatic exclusion rule.

Table P3.4

Descriptive reductions by department

Department	n	Mean reduction	Median
Logistics	4	2.75	3.0
Production	8	1.88	2.0
QA	8	0.62	1.5

Note. Department summaries are exploratory because subgroup sizes are small.

The outlier screen and department summaries support interpretation rather than hypothesis testing. If a unit is unusual, the correct next step is to check records and implementation notes. If departments differ descriptively, that should guide future sampling or process review rather than a formal subgroup claim.

Reporting Summary

Incident counts decreased after the process change, with a mean after-before difference of -1.55, 95% CI [-2.60, -0.50], p = 0.006. The signed-rank sensitivity analysis led to the same substantive conclusion. The finding should be framed as strong pilot evidence of improvement, with the usual caution that department-level patterns remain exploratory.

Extension Task

Repeat the paired analysis after excluding the largest absolute reduction as a sensitivity check. Report whether the mean difference, confidence interval and signed-rank conclusion change enough to affect the practical interpretation.

Project 4. Evaluating a Reading Intervention in Small Classrooms (Education)

Background

A school piloted a reading intervention with 22 students across two grades and two teachers. Each student completed a pre-test and post-test. The paired design is appropriate because each student serves as their own baseline, but the classroom context still limits generalisability.

Descriptive Summary

Table P4.1

Reading-score summary for the classroom pilot

Quantity	Value
Students	22
Mean pre-test score	65.6
Mean post-test score	73.1
Mean improvement	7.6
Median improvement	8.0
SD of improvement	3.7

Note. Improvement is post-test minus pre-test; positive values favour the intervention period.

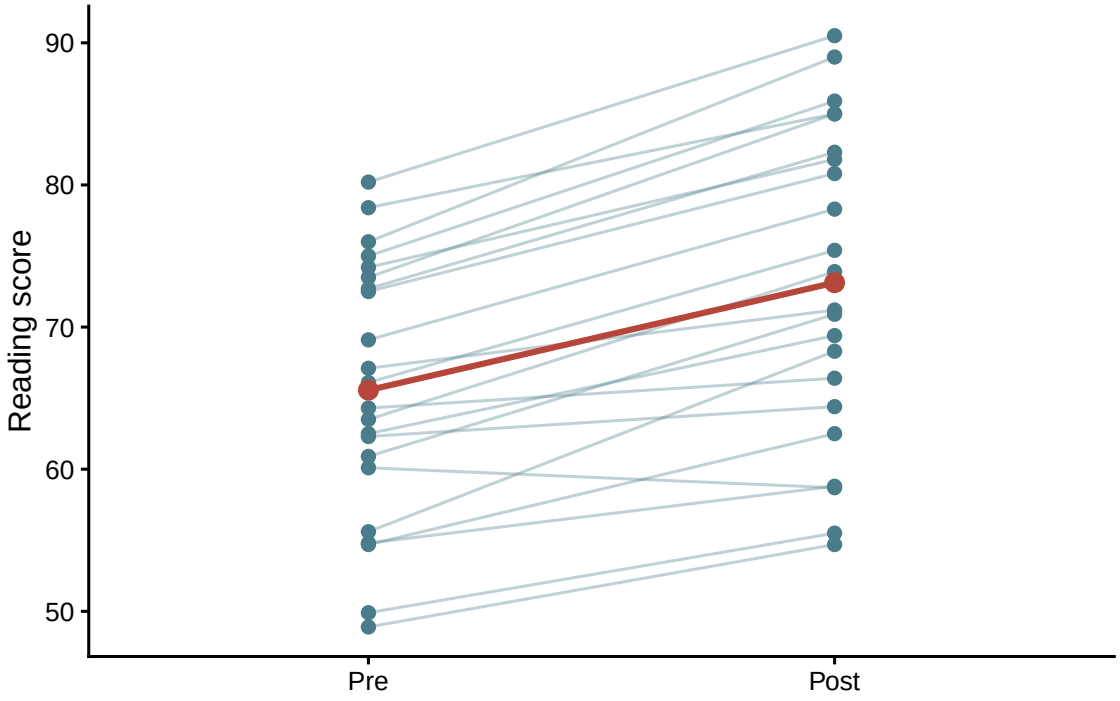


Figure P4.1: Student reading scores before and after the classroom intervention.

Most student trajectories increase from pre-test to post-test. The connected-line figure is useful because it shows both the average gain and the variability in individual responses.

Primary Analysis

Table P4.2

Paired reading-score analysis

Analysis	Estimate	95% CI	p-value
Paired t-test	Mean improvement = 7.56	5.92 to 9.20	<0.001
Wilcoxon signed-rank	Pseudomedian improvement = 7.80	5.90 to 9.40	<0.001
Shapiro-Wilk on improvements	W = 0.96	–	0.432
Paired standardised mean change	dz = 2.04	Not computed	–

Note. The signed-rank test is included as a sensitivity check. The Shapiro-Wilk test is reported descriptively because normality tests have limited power in small samples.

The estimated mean improvement is about 7.6 points, and both the t-test and signed-rank sensitivity analysis support a positive gain. The paired standardised change is large, but it should be interpreted as pilot evidence from a specific school context rather than as a stable population effect.

Distribution and Classroom Context

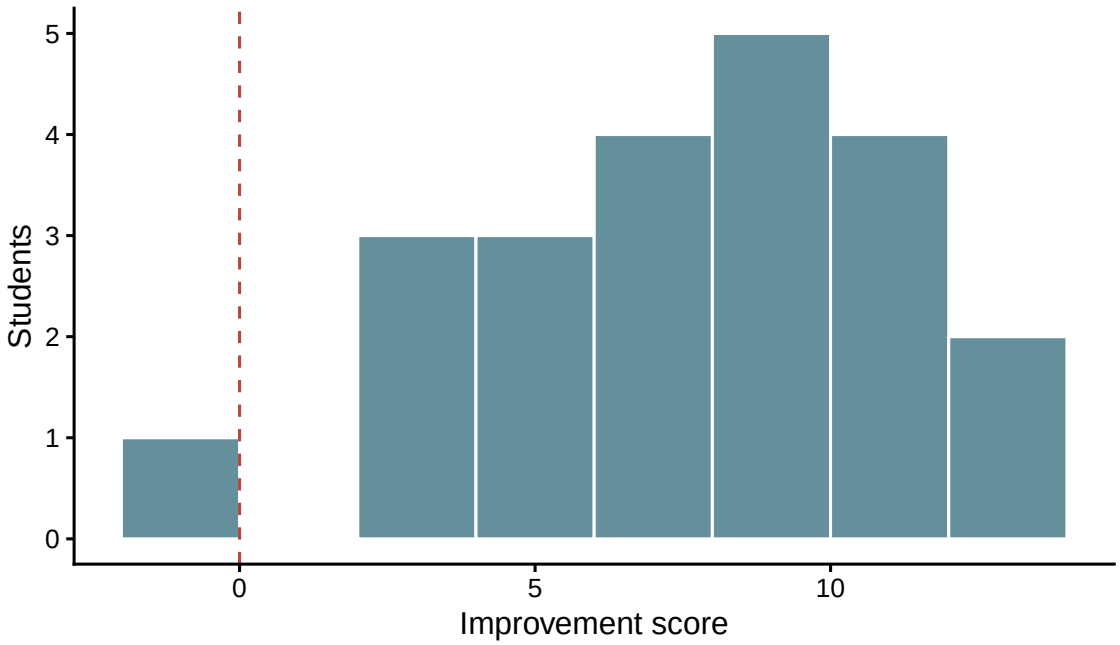


Figure P4.2: Distribution of individual reading-score improvements.

Table P4.3

Descriptive improvements by grade and teacher

Grade	Teacher	n	Mean improvement	Median
3rd	Mr. Lee	4	8.93	10.5
3rd	Ms. Johnson	2	9.85	9.9
4th	Mr. Lee	7	6.41	8.3
4th	Ms. Johnson	9	7.34	6.9

Note. These cells are too small for formal teacher or grade comparisons; they are included to support contextual interpretation.

The classroom summaries help readers judge whether the overall gain appears concentrated in one grade or teacher. Because the cells are small and not independently randomised, the table should be treated as implementation context rather than evidence of differential effectiveness.

Reporting Summary

Reading scores increased from pre-test to post-test by a mean of 7.56 points, 95% CI [5.92, 9.20], $p < 0.001$. The result is promising for the participating classrooms. A stronger study would include a comparison group, prespecified fidelity checks, and a larger sample of classrooms.

Extension Task

Create a forest plot of mean improvement by grade or teacher using the descriptive summaries in Table P4.3. Add a note explaining why the plot is useful for implementation review but not sufficient for formal subgroup inference.

Project 5. Understanding Intervention Mechanisms: Simple Mediation Analysis

Background

This project examines whether a study-skills intervention is associated with exam scores partly through self-efficacy. The worked example uses a balanced n = 70 subset, with 35 students per group, to keep the project within the small-sample scope of the book. Mediation language requires care. Even with a randomised intervention, the mediator is usually not randomised, so the indirect path can be confounded by unmeasured variables. The results should therefore be described as consistent with a mediation pattern, not as proof of a causal mechanism.

Data Summary

Table P5.1

Self-efficacy and exam scores by intervention group

Group	n	Mean self-efficacy	SD self-efficacy	Mean exam score	SD exam score
Control	35	4.22	0.28	65.20	2.10
Intervention	35	6.89	0.34	81.17	2.71

Note. The intervention indicator is coded 1 for intervention and 0 for control in the regression models.

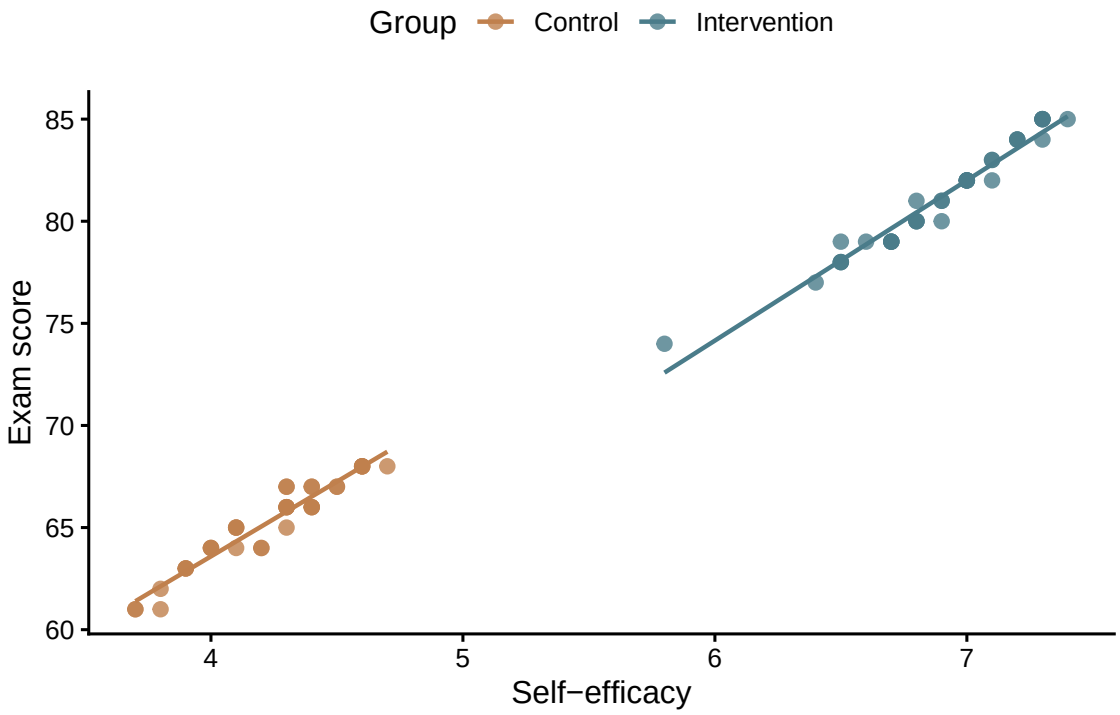


Figure P5.1: Association between self-efficacy and exam score by intervention group.

The intervention group has higher self-efficacy and higher exam scores. The scatterplot also shows a strong association between self-efficacy and exam performance, which motivates the mediation model.

Regression Paths

Table P5.2

Regression paths for the simple mediation model

Path	Estimate	95% CI	p-value
c: Intervention -> exam score	15.97	14.82 to 17.13	<0.001
a: Intervention -> self-efficacy	2.67	2.53 to 2.82	<0.001
b: Self-efficacy -> exam score | intervention	7.64	7.19 to 8.09	<0.001
c': Intervention -> exam score | self-efficacy	-4.45	-5.69 to -3.21	<0.001

Note. Path c is the total intervention association with exam score. Path c' is the direct association after adjusting for self-efficacy.

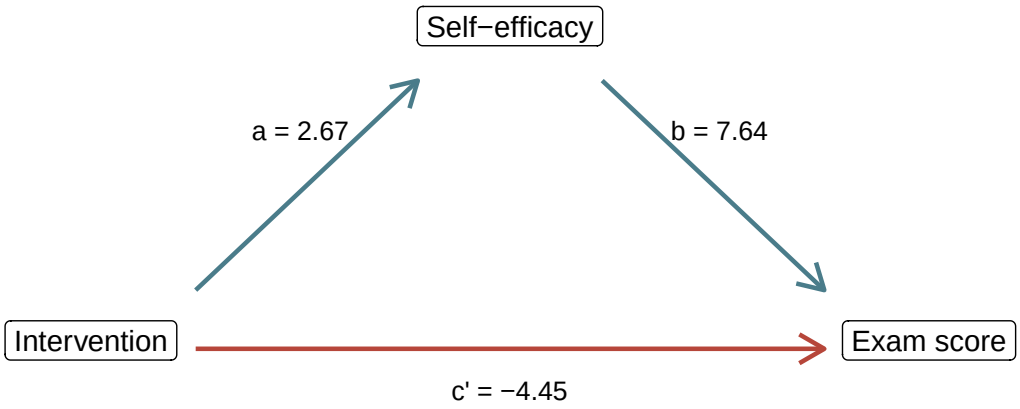


Figure P5.2: Estimated paths in the simple mediation model.

The path pattern is strong but not simple. The indirect path is large because the intervention is strongly associated with self-efficacy and self-efficacy is strongly associated with exam score. After adjustment for self-efficacy, the direct intervention coefficient reverses direction. That reversal is a warning to interpret the model as a statistical decomposition rather than as a clean causal story.

Bootstrap Indirect Effect

Table P5.3

Bootstrap summary of the indirect effect

Quantity	Value
Total effect (c)	15.97
Indirect effect (a x b)	20.42
Bootstrap 95% CI for indirect effect	18.36 to 22.75
Direct effect (c')	-4.45
Bootstrap resamples	5000

Note. The bootstrap resamples rows with replacement using set.seed(2025). The interval describes uncertainty in the product a x b under the fitted linear models.

The bootstrap interval for the indirect effect is well above zero. The approach used here — multiplying path coefficients and bootstrapping their product — is equivalent to Hayes PROCESS Model 4 under a basic percentile bootstrap. Bias-corrected and accelerated (BCa) bootstrap intervals, available via the `mediation` package, are generally preferred for small samples because they adjust for bootstrap distribution skewness. The result is consistent with self-efficacy carrying

a substantial portion of the intervention association with exam scores. However, because the direct effect reverses after adjustment, the analysis should be reported as inconsistent or suppressor-like mediation rather than as a straightforward partial-mediation story. Suppression occurs when adjustment for a mediator or related variable changes the sign or magnitude of the direct path so that the indirect effect can exceed the total effect (MacKinnon, Krull, and Lockwood 2000).

Reporting Summary

The intervention was associated with higher exam scores in the total-effect model, b = 15.97. It was also associated with higher self-efficacy, a = 2.67, and self-efficacy was strongly associated with exam score after adjustment, b = 7.64. The bootstrap indirect effect was 20.42, 95% CI [18.36, 22.75]. These results are consistent with an indirect pathway through self-efficacy, but the direct-effect reversal is a suppressor-like pattern (MacKinnon, Krull, and Lockwood 2000). The mechanism should be presented cautiously and checked in a design with temporal separation and stronger confounding control.

Extension Task

Re-run the project with a different reproducible subset size, such as 25 or 45 students per group, and compare the indirect-effect interval with the n = 70 analysis. Then add one sensitivity paragraph explaining whether the mediation interpretation changes when the available sample changes.

Summary of Part E

Part E brought the earlier material together through complete worked projects. The projects showed how to combine design decisions, estimation, nonparametric methods, reliability analysis, paired comparisons, subgroup exploration, mediation analysis, tables, figures, and transparent interpretation in realistic small-sample applications across business, service quality, operations, and education contexts.

The main value of the worked projects is integration. Rather than treating each method in isolation, they show how a defensible small-sample analysis is assembled from linked choices about the question, the data, the method, the sensitivity checks and the reporting. The common reporting pattern is deliberately cautious: state the estimand, show the data structure, report uncertainty, use sensitivity analyses, and avoid causal or subgroup claims that the design cannot support.

References

Ayre, Colin, and Andrew J. Scally. 2014. 'Critical Values for Lawshe's Content Validity Ratio: Revisiting the Original Methods of Calculation'. *Measurement and Evaluation in Counseling and Development* 47 (1): 79–86. https://doi.org/10.1177/0748175613513808.

Briggs, Steven R., and Jonathan M. Cheek. 1986. 'The Role of Factor Analysis in the Development and Evaluation of Personality Scales'. *Journal of Personality* 54 (1): 106–48.

Buuren, Stef van. 2018. *Flexible Imputation of Missing Data*. 2nd ed. Boca Raton, FL: Chapman; Hall/CRC.

Cameron, A. Colin, and Pravin K. Trivedi. 2013. *Regression Analysis of Count Data*. 2nd ed. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139013567.

Chen, Henian, Patricia Cohen, and Sophie Chen. 2010. 'How Big Is a Big Odds Ratio? Interpreting the Magnitudes of Odds Ratios in Epidemiological Studies'. *Communications in Statistics - Simulation and Computation* 39 (4): 860–64. https://doi.org/10.1080/03610911003650383.

Cicchetti, Domenic V. 1994. 'Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology'. *Psychological Assessment* 6 (4): 284–90. https://doi.org/10.1037/1040-3590.6.4.284.

Clark, Lee Anna, and David Watson. 1995. 'Constructing Validity: Basic Issues in Objective Scale Development'. *Psychological Assessment* 7 (3): 309–19.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates.

Conover, W. J. 1999. *Practical Nonparametric Statistics*. 3rd ed. New York: John Wiley & Sons.

Costello, Anna B., and Jason Osborne. 2005. 'Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most from Your Analysis'. *Practical Assessment, Research, and Evaluation* 10 (1): 1–9. https://doi.org/10.7275/jyj1-4868.

Davison, A. C., and D. V. Hinkley. 1997. *Bootstrap Methods and Their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

DeVellis, Robert F., and Carolyn T. Thorpe. 2021. *Scale Development: Theory and Applications*. 5th ed. Thousand Oaks, CA: SAGE Publications.

Dienes, Zoltan. 2014. 'Using Bayes to Get the Most Out of Non-Significant Results'. *Frontiers in Psychology* 5: 781. https://doi.org/10.3389/fpsyg.2014.00781.

Efron, Bradley, and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York, NY: Chapman; Hall/CRC.

Elm, Erik von, Douglas G. Altman, Matthias Egger, Stuart J. Pocock, Peter C. Gøtzsche, and Jan P. Vandenbroucke. 2007. 'The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies'. *PLoS Medicine* 4 (10): e296. https://doi.org/10.1371/journal.pmed.0040296.

Fay, Michael P. 2010. 'Confidence Intervals That Match Fisher's Exact or Blaker's Exact Tests'. *Biostatistics* 11 (2): 373–74. https://doi.org/10.1093/biostatistics/kxp050.

Firth, David. 1993. 'Bias Reduction of Maximum Likelihood Estimates'. *Biometrika* 80 (1): 27–38. https://doi.org/10.1093/biomet/80.1.27.

Gelman, Andrew, Daniel Simpson, and Michael Betancourt. 2017. 'The Prior Can Often Only Be Understood in the Context of the Likelihood'. *Entropy* 19 (10): 555. https://doi.org/10.3390/e19100555.

Good, Phillip I. 2005. *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. 3rd ed. New York: Springer.

Graham, John W. 2009. 'Missing Data Analysis: Making It Work in the Real World'. *Annual Review of Psychology* 60: 549–76. https://doi.org/10.1146/annurev.psych.58.110405.085530.

Hair, Joseph F., G. Tomas M. Hult, Christian M. Ringle, and Marko Sarstedt. 2017. *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*. 2nd ed. Thousand Oaks, CA: SAGE Publications.

Harrell, Frank E. 2015. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. 2nd ed. Springer Series in Statistics. New York: Springer. https://doi.org/10.1007/978-3-319-19425-7.

Heinze, Georg, and Michael Schemper. 2002. 'A Solution to the Problem of Separation in Logistic Regression'. *Statistics in Medicine* 21 (16): 2409–19. https://doi.org/10.1002/sim.1047.

Hodges, Jr., J. L., and E. L. Lehmann. 1963. 'Estimates of Location Based on Rank Tests'. *The Annals of Mathematical Statistics* 34 (2): 598–611. https://doi.org/10.1214/aoms/1177704172.

Hoenig, John M., and Dennis M. Heisey. 2001. 'The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis'. *The American Statistician* 55 (1): 19–24.

Hosmer, David W., Stanley Lemeshow, and Rodney X. Sturdivant. 2013. *Applied Logistic Regression*. 3rd ed. Wiley Series in Probability and Statistics. Hoboken, NJ: John Wiley & Sons. https://doi.org/10.1002/9781118548387.

Hu, Li-tze, and Peter M. Bentler. 1999. 'Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives'. *Structural Equation Modeling* 6 (1): 1–55. https://doi.org/10.1080/10705519909540118.

Huberty, Carl J., and Stephen Olejnik. 2006. *Applied MANOVA and Discriminant Analysis*. 2nd ed. Hoboken, NJ: Wiley. https://doi.org/10.1002/047178947X.

Hwang, Ching-Lai, and Kwangsun Yoon. 1981. *Multiple Attribute Decision Making: Methods and Applications*. Berlin: Springer. https://doi.org/10.1007/978-3-642-48318-9.

Hyndman, Rob J., and George Athanasopoulos. 2021. *Forecasting: Principles and Practice*. 3rd ed. Melbourne, Australia: OTexts. https://otexts.com/fpp3/.

Koo, Terry K., and Mae Y. Li. 2016. 'A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research'. *Journal of Chiropractic Medicine* 15 (2): 155–63. https://doi.org/10.1016/j.jcm.2016.02.012.

Kraft, Matthew A. 2020. 'Interpreting Effect Sizes of Education Interventions'. *Educational Researcher* 49 (4): 241–53. https://doi.org/10.3102/0013189X20912798.

Lakens, Daniël. 2013. ‘Calculating and Reporting Effect Sizes to Facilitate Cumulative Science: A Practical Primer for t-Tests and ANOVAs’. *Frontiers in Psychology* 4: 863. https://doi.org/10.3389/fpsyg.2013.00863.

Lakens, Daniël, Anne M. Scheel, and Peder M. Isager. 2018. ‘Equivalence Testing for Psychological Research: A Tutorial’. *Advances in Methods and Practices in Psychological Science* 1 (2): 259–69. https://doi.org/10.1177/2515245918770963.

Lawshe, Charles H. 1975. ‘A Quantitative Approach to Content Validity’. *Personnel Psychology* 28 (4): 563–75. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x.

Lipsey, Mark W., and David B. Wilson. 1993. ‘The Efficacy of Psychological, Educational, and Behavioural Treatment: Confirmation from Meta-Analysis’. *American Psychologist* 48 (12): 1181–1209. https://doi.org/10.1037/0003-066X.48.12.1181.

Little, Roderick J. A. 1988. ‘A Test of Missing Completely at Random for Multivariate Data with Missing Values’. *Journal of the American Statistical Association* 83 (404): 1198–1202. https://doi.org/10.1080/01621459.1988.10478722.

MacKinnon, David P., Jennifer L. Krull, and Chondra M. Lockwood. 2000. ‘Equivalence of the Mediation, Confounding and Suppression Effect’. *Prevention Science* 1 (4): 173–81.

Mair, Patrick, and Rand R. Wilcox. 2020. ‘Robust Statistical Methods in r Using the WRS2 Package’. *Behavior Research Methods* 52 (2): 464–88. https://doi.org/10.3758/s13428-019-01246-w.

Mann, Henry B., and Donald R. Whitney. 1947. ‘On a Test of Whether One of Two Random Variables Is Stochastically Larger Than the Other’. *The Annals of Mathematical Statistics* 18 (1): 50–60. https://doi.org/10.1214/aoms/1177730491.

McDonald, Roderick P. 1999. *Test Theory: A Unified Treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Morey, Richard D., and Jeffrey N. Rouder. 2011. ‘Bayes Factor Approaches for Testing Interval Null Hypotheses’. *Psychological Methods* 16 (4): 406–19. https://doi.org/10.1037/a0024377.

Nielsen, Jakob. 1993. *Usability Engineering*. San Francisco, CA: Morgan Kaufmann.

Olsson, Ulf. 1979. ‘Maximum Likelihood Estimation of the Polychoric Correlation Coefficient’. *Psychometrika* 44 (4): 443–60.

Opricovic, Serafim, and Gwo-Hshiung Tzeng. 2004. ‘Compromise Solution by MCDM Methods: A Comparative Analysis of VIKOR and TOPSIS’. *European Journal of Operational Research* 156 (2): 445–55. https://doi.org/10.1016/S0377-2217(03)00020-1.

Page, Matthew J., Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, et al. 2021. ‘The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews’. *BMJ* 372: n71. https://doi.org/10.1136/bmj.n71.

R Core Team. 2026. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York, NY: Wiley.

Saaty, Thomas L. 1980. *The Analytic Hierarchy Process*. New York, NY: McGraw-Hill.

Schulz, Kenneth F., Douglas G. Altman, and David Moher. 2010. ‘CONSORT 2010 Statement: Updated Guidelines for Reporting Parallel Group Randomised Trials’. *BMJ* 340: c332. https:

//doi.org/10.1136/bmj.c332.

Shan, Guogen. 2018. *Exact Statistical Inference for Categorical Data*. London: Academic Press.

Sterne, Jonathan A. C., Ian R. White, John B. Carlin, Michael Spratt, Patrick Royston, Michael G. Kenward, Angela M. Wood, and James R. Carpenter. 2009. 'Multiple Imputation for Missing Data in Epidemiological and Clinical Research: Potential and Pitfalls'. *BMJ* 338: b2393. https://doi.org/10.1136/bmj.b2393.

Teare, M. Dawn, Munyaradzi Dimairo, Neil Shephard, Amanda Hayman, Amy Whitehead, and Stephen J. Walters. 2014. 'Sample Size Requirements to Estimate Key Design Parameters from External Pilot Randomised Controlled Trials: A Simulation Study'. *Trials* 15: 264. https://doi.org/10.1186/1745-6215-15-264.

Tomczak, Maciej, and Ewa Tomczak. 2014. 'The Need to Report Effect Size Estimates Revisited: An Overview of Some Recommended Measures of Effect Size'. *Trends in Sport Sciences* 21 (1): 19–25.

Trizano-Hermosilla, Italo, and Jose M. Alvarado. 2016. 'Best Alternatives to Cronbach's Alpha Reliability in Realistic Conditions: Congeneric and Asymmetrical Measurements'. *Frontiers in Psychology* 7: 769. https://doi.org/10.3389/fpsyg.2016.00769.

Van de Schoot, Rens, and Milica Miočević. 2020. *Small Sample Size Solutions: A Guide for Applied Researchers and Practitioners*. London: Routledge.

Vehtari, Aki, Andrew Gelman, and Jonah Gabry. 2017. 'Practical Bayesian Model Evaluation Using Leave-One-Out Cross-Validation and WAIC'. *Statistics and Computing* 27 (5): 1413–32. https://doi.org/10.1007/s11222-016-9696-4.

Wagenmakers, Eric-Jan, Jonathon Love, Maarten Marsman, Tahira Jamil, Alexander Ly, Josine Verhagen, Ravi Selker, et al. 2018. 'Bayesian Inference for Psychology. Part II: Example Applications with JASP'. *Psychonomic Bulletin & Review* 25 (1): 58–76. https://doi.org/10.3758/s13423-017-1323-7.

White, Ian R., Patrick Royston, and Angela M. Wood. 2011. 'Multiple Imputation Using Chained Equations: Issues and Guidance for Practice'. *Statistics in Medicine* 30 (4): 377–99. https://doi.org/10.1002/sim.4067.